Distributionally Robust Constrained Reinforcement Learning under Strong Duality

Zhengfei Zhang*
zhengfei.zhang@sc.tsinghua.edu.cn
Tsinghua University

Kishan Panaganti
kpb@caltech.edu
California Institute of Technology

Laixi Shi
laixis@caltech.edu
California Institute of Technology

Yanan Sui
ysui@tsinghua.edu.cn
Tsinghua University

Adam Wierman
adamw@caltech.edu
California Institute of Technology

Yisong Yue
yyue@caltech.edu
California Institute of Technology

Abstract

We study the problem of Distributionally Robust Constrained RL (DRC-RL), where the goal is to maximize the expected reward subject to environmental distribution shifts and constraints. This setting captures situations where training and testing environments differ, and policies must satisfy constraints motivated by safety or limited budgets. Despite significant progress toward algorithm design for the separate problems of distributionally robust RL and constrained RL, there do not yet exist algorithms with end-to-end convergence guarantees for DRC-RL. We develop an algorithmic framework based on strong duality that enables the first efficient and provable solution in a class of environmental uncertainties. Further, our framework exposes an inherent structure of DRC-RL that arises from the combination of distributional robustness and constraints, which prevents a popular class of iterative methods from tractably solving DRC-RL, despite such frameworks being applicable for each of distributionally robust RL and constrained RL individually. Finally, we conduct experiments on a car racing benchmark to evaluate the effectiveness of the proposed algorithm.

1 Introduction

In many real-world decision-making tasks, policies must not only be reward-maximizing but also be robust to environmental distribution shifts while satisfying application constraints. Environmental distribution shifts occur in scenarios where there is a mismatch between the training and testing environments, such as due to environment changes (Maraun, 2016), modeling errors (Chen et al., 1996), or adversarial disturbances (Pioch et al., 2009). Constraints are imposed in tasks that require adherence to safety factors (Haddadin et al., 2012; Weidemann et al., 2023), budgets in strategy games (Vinyals et al., 2019), diverse interests in advertisement recommendations (KRM et al., 2021; Bagenal et al., 2023), and so on. This motivates us to tackle both challenges simultaneously, inspiring the study of problems called distributionally robust constrained RL (DRC-RL) (Russel et al., 2020; Wang et al., 2022).

*This work was initiated during the visiting undergraduate research program at the California Institute of Technology.

The goal of DRC-RL is to learn a policy that simultaneously optimizes the expected reward and satisfies the constraints in the worst-case scenario when the deployed environment deviates from the nominal one within a prescribed uncertainty set. DRC-RL has received growing attention in recent years and is typically modeled as a constrained optimization problem. As it is unknown if the strong duality holds for the DRC-RL problems, most recent works either use different formulations (e.g., risk-averse) to consider environmental uncertainty (Queeney & Benosman, 2024; Kim et al., 2024), or simply focus on one of the primal (Sun et al., 2024) and the dual problem (Wang et al., 2022; Bossens, 2023), such that an end-to-end guarantee is still absent. More broadly, there has been significant progress in developing rigorous algorithms that address the two challenges that make up DRC-RL individually: distributionally robust RL (DR-RL) (Iyengar, 2005; Wiesemann et al., 2013; Li et al., 2022; Panaganti et al., 2022) and constrained RL (C-RL) (Le et al., 2019; Miryoosefi et al., 2019; Efroni et al., 2020; Ding et al., 2021). Many of these works have focused on a simple, intuitive, greedy policy induced by taking the greedy (best) action with respect to the current learned value functions. This raises the question of whether a similar greedy approach can be effective in DRC-RL or if additional challenges arise from the combination of distributional robustness and constraints.

To address this question, in this paper, we develop a general framework that transfers the policy learning problem to a game-theoretic formulation with a constructed strong duality, where the dual problem is treated as a player's objective. In DR-RL and C-RL, targets similar to this dual function are solved via greedy policies (Iyengar, 2005; Le et al., 2019). Mathematically, one can think of such a procedure as applying an operator efficiently and greedily, and convergence depends on proving that this operator is a contraction. While such a greedy approach works for DR-RL (Iyengar, 2005) and C-RL (Miryoosefi et al., 2019) in isolation, we show that only with further assumptions can one apply this approach to DRC-RL, e.g., for R-contamination uncertainty sets (Huber, 1965; Wang & Zou, 2022). We prove that, in general, no such operator exists for the joint DRC-RL problem, implying an impossibility result for a commonly applied class of algorithms.

In summary, this paper makes the following main contributions:

- We propose a multi-level systematic framework to solve DRC-RL for general uncertainty sets in Section 3. We show that guarantees for subroutines combine to ensure end-to-end guarantees for DRC-RL.
- Focusing on the R-contamination uncertainty set, we instantiate our framework to provide the first provable efficient solution for DRC-RL in Section 4. Our solution uses a shortened horizon in subroutines to ensure distributional robustness. We verify its effectiveness with an experiment using a high-dimensional car-racing task.
- We consider general uncertainty sets in Section 5 and show that the combination of constraints and distributional robustness requirements yields that DRC-RL cannot be solved by considering greedy policies, which is the key of a popular class of iteration methods proposed previously for standard RL, DR-RL, and C-RL problems (Iyengar, 2005; Le et al., 2019).

Notation. For any set $\mathcal{S}$, $\Delta(\mathcal{S})$ denotes the set of probability distribution over $\mathcal{S}$. We use $\otimes_i \mathcal{X}_i$ to denote a product space of spaces $\mathcal{X}_i$'s. We use $\langle x, y\rangle$ to denote the inner product of two same-sized vectors. We claim $f(T)$ is $o(T)$ if $f(T)/T \to 0$ as $T \to \infty$. $f(x) \leq O(g(x))$ denote $f(x) \leq Cg(x)$ for some positive constant C.

2 Preliminaries and Problem Formulation

Robust Markov Decision Process. A robust Markov decision process (MDP) with infinite horizon can be specified by a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma, \mu)$, where $\mathcal{S}$ denotes the finite state space[1], $\mathcal{A}$ denotes the finite action space, $r : \mathcal{S} \times \mathcal{A} \to [0, \bar{R}]$ is the known deterministic reward function with some positive maximal magnitude $\bar{R}$, $\gamma \in [0, 1)$ is the discount factor, and μ is the initial

[1]We study the finite state space here for simplicity sake, while all theoretical and empirical results in the main text should hold for any bounded and closed finite-dimensional state space as well.

state distribution. In this paper, we focus on (s, a)-rectangularity uncertainty set for the transition kernel $\mathcal{P}$ (Nilim & Ghaoui, 2003; Iyengar, 2005), i.e. $\mathcal{P} = \otimes_{(s,a)\in\mathcal{S}\times\mathcal{A}}\mathcal{P}_{s,a}$, where $\mathcal{P}_{s,a} = \{P_{s,a} \in \Delta(\mathcal{S})|D(P_{s,a}, P^o_{s,a}) \leq \beta_{s,a}\}$, describing a neighborhood of the nominal model $P^o = (P^o_{s,a}, (s, a) \in \mathcal{S} \times \mathcal{A})$ by some measurement function $D : \Delta(\mathcal{S}) \times \Delta(\mathcal{S}) \to \mathbb{R}$ and robustness level $\beta \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$. We consider any policy $\pi : \mathcal{S} \to \mathcal{A}$ in the class of deterministic policies Π. The robust value function of a policy π is then defined as the worst-case accumulated reward following the policy π over MDPs within the uncertainty set as below (Iyengar, 2005, Sec.3):

$$V^\pi_r(s) := \min_{\mathcal{K}\in\otimes_{t\geq 0}\mathcal{P}} \mathbb{E}_{\mathcal{K}}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)|s_0 = s, \pi]. \tag{1}$$

Here $\mathcal{K}$ is a sequence of transition kernels within the same uncertainty set $\mathcal{P}$ over each time step. Moreover, for any policy π, the robust value function V^π_r is the unique stationary point of the robust Bellman consistency operator (Iyengar, 2005),

$$\mathcal{T}^\pi_{r,\text{rob}} v(s) := r(s, \pi(s)) + \gamma \min_{P\in\mathcal{P}} \langle P_{s,\pi(s)}, v\rangle. \tag{2}$$

The ultimate goal of distributionally robust RL is to find the optimal robust policy π^* that attains maximized robust value function $V^*_r = \max_\pi V^\pi_r$. To attain the optimal robust value V^*_r, we have various dynamic programming procedures (Iyengar, 2005), e.g. iterations $v_{k+1} = \mathcal{T}^*_{r,\text{rob}} v_k$ converge to V^*_r, where the optimality operator $\mathcal{T}^*_{r,\text{rob}} v(s) := \max_\pi \mathcal{T}^\pi_{r,\text{rob}} v(s) = \mathcal{T}^{\mathcal{G}(v)}_{r,\text{rob}} v(s)$ using the greedy policy $\mathcal{G}(v)[s] := \arg\max_a\{r(s, a) + \gamma \min_{P\in\mathcal{P}}\langle P_{s,a}, v\rangle\}$. Typically, such problems that are usually solved by greedy policy $\mathcal{G}(v)[s] = \arg\max_\pi \mathcal{T}^\pi_{r,\text{rob}} v(s)$ has the name of policy improvement step.

Distributionally Robust Constrained RL (DRC-RL). We formulate the distributionally robust constrained MDP as a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r, g, \tau, \gamma, \mu)$, where $\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma, \mu$ are identical to that in robust MDPs. Here, $g := [g_1, g_2, \cdots g_m]$ with $g_i : \mathcal{S} \times \mathcal{A} \to [0, \tau_i]$ for all $i \in 1, 2, \cdots, m$, representing the aggregation vector of m known deterministic reward-based constraint functions based on the constraint thresholds $\tau = [\tau_1, \cdots, \tau_m]$. We aim to learn a policy π within the deterministic policy class denoted as Π. As environmental distribution shifts can be applied to constraints and the objective independently, e.g. estimation errors, we formulate the goal of distributionally robust constrained RL (DRC-RL) as solving the following constrained optimization problem:

$$\max_{\pi\in\Pi} V^\pi_r(\mu) \;\; \text{s.t.} \;\; V^\pi_{g_i}(\mu) \geq \tau_i, 1 \leq i \leq m, \tag{3}$$

where $V^\pi_r, V^\pi_{g_i}$ is the robust value functions (1) corresponding to the objective reward function r and constraint functions g_i's, and their corresponding expected robust values according to the initial state distributions are $V^\pi_r(\mu) = \langle V^\pi_r, \mu\rangle$ and $V^\pi_{g_i}(\mu) = \langle V^\pi_{g_i}, \mu\rangle$. For brevity sake, we denote the constraint vector as $V^\pi_g := [V^\pi_{g_1}, V^\pi_{g_2}, \cdots, V^\pi_{g_m}]^\top \in \mathbb{R}^m$.

3 DRC-RL with General Uncertainty Sets

In this section, we develop a general framework and meta algorithm for DRC-RL with an arbitrary uncertainty set in Section 3.1 and introduce the subroutines of the framework in Section 3.2.

3.1 A Meta Algorithm for DRC-RL

Constrained RL can be viewed as a constrained optimization problem that has been proven to have strong duality generally by Paternain et al. (2019). However, it is currently not known whether DRC-RL maintains strong duality. To show strong duality of DRC-RL problem (3), we consider a class of mixed policies denoted as $Conv(\Pi)$, defined as below (Miryoosefi et al., 2019; Le et al., 2019):

$$\left\{\pi_{\alpha,\{\pi_i\}_{i=1}^T} \sim \textsf{Categorical}\left(\{\pi_i\}_{i=1}^T, \alpha\right) : 0 < T < \infty, \pi_i \in \Pi, \forall i; \alpha = [\alpha_1, \cdots, \alpha_T] \in \Delta(T)\right\}, \tag{4}$$

Algorithm 1 Meta Algorithm for DRC-RL

for each round t **do**
  $\pi_t \leftarrow \text{BestResponse}(\lambda_t)$ ▷ Non-trivial for DRC-RL problems
  $\hat{\pi}_t \leftarrow \frac{1}{t}\sum_{t'=1}^{t} \pi_{t'}, \hat{\lambda}_t \leftarrow \frac{1}{t}\sum_{t'=1}^{t} \lambda_{t'}$ ▷ Mixed policy $\hat{\pi}_t$
  $L_{\max} = L(\text{BestResponse}(\hat{\lambda}_t), \hat{\lambda}_t)$
  $L_{\min} = \min_{\lambda} L(\hat{\pi}_t, \lambda)$
  if $L_{\max} - L_{\min} < \omega$ **then** ▷ Calculating current duality gap
    return $\hat{\pi}_t$
  end if
  $\lambda_{t+1} \leftarrow \text{OnlineAlgo}(\pi_1, ..., \pi_{t-1}, \pi_t)$
end for

where $\mathsf{Categorical}(\{\pi_i\}_{i=1}^T, \alpha)$ is a categorical distribution such that $\mathbb{P}(\pi_{\alpha,\{\pi_i\}_{i=1}^T} = \pi_i) = \alpha_i$ for all $i = 1, 2, \cdots, T$. To execute any mixed policy $\pi_{\alpha,\{\pi_i\}_{i=1}^T} \in Conv(\Pi)$, at the beginning of each episode, a deterministic policy π is sampled independently from $\mathsf{Categorical}(\{\pi_i\}_{i=1}^T, \alpha)$ and serve as the action selection rule for the entire episode. Thus, the robust value function of a mixed policy $V_r^{\pi_{\alpha,\{\pi_i\}_{i=1}^T}}$ is defined as $V_r^{\pi_{\alpha,\{\pi_i\}_{i=1}^T}} := \mathbb{E}_{\pi\sim\mathsf{Categorical}(\{\pi_i\}_{i=1}^T,\alpha)}[V_r^{\pi}] = \sum_{i=1}^T \alpha_i V_r^{\pi_i}$.

Proposition 3.1. *When substituting Π with its convex hull $Conv(\Pi)$ in the DRC-RL problem (3), strong duality holds if Slater's condition holds.*

The proof of the above proposition and other results of this section are postponed to Appendix A.

We assume the DRC-RL problem (3) is feasible and that Slater's condition (Boyd & Vandenberghe, 2004) holds, where the latter only requires the existence of an interior solution upon feasibility. The problem considering the augmented solution class $Conv(\Pi)$ has a solution no worse than the original problem (3), and the convexification itself does not pose any restriction on deterministic policies. As such, we directly denote the convex hull $Conv(\Pi)$ as Π in the rest of the paper, and always treat π as a mixed policy unless specified.

The Lagrangian of (3) is $L(\pi, \lambda) := V_r^{\pi}(\mu) - \lambda^{\top}(V_g^{\pi}(\mu) - \tau)$ for some $\lambda \in \mathbb{R}_+^m$. Strong duality indicates $\max_{\pi\in\Pi} \min_{\lambda\in\mathbb{R}_+^m} L(\pi, \lambda) = \min_{\lambda\in\mathbb{R}_+^m} \max_{\pi\in\Pi} L(\pi, \lambda)$. By the definition of mixed polices, $V_r^{\pi}(\mu)$ and $V_g^{\pi}(\mu)$ are all linear to policy π (see Appendix A.1). Therefore, $L(\pi, \lambda)$ is linear to both λ and π, and a game-theoretic perspective can be applied. That is, we view the problem as a two-player game between a π-player and a λ-player (Freund & Schapire, 1999; Miryoosefi et al., 2019).

Algorithm 1 describes this repeated game, where both players seek to decrease the duality gap. The π-player runs `Best-response` to maximize Lagrangian $L(\pi, \lambda_t)$ given the current λ_t,

$$\pi_t := \text{Best-response}(\lambda_t) \in \text{argmax}_{\pi} L(\pi, \lambda_t). \tag{5}$$

The λ-player then employs any no-regret `Online-Algorithm` (Shalev-Shwartz, 2007) to minimize $L(\pi_t, \lambda)$, which satisfies:

$$\sum_t (-L)(\pi_t, \lambda_t) \geq \max_{\lambda} \sum_t (-L)(\pi_t, \lambda) - o(T). \tag{6}$$

Algorithm 1 terminates when the estimated primal-dual gap is below a threshold w.

Proposition 3.2. *Algorithm 1 is guaranteed to converge if (i) `Best-response` gives the best deterministic policy in the deterministic policy class (ii) $L_{\max}$ and $L_{\min}$ in Algorithm 1 are precisely evaluated. Additionally, the exact convergence rate depends on the regret of `Online-Algorithm`.*

3.2 The `Online-Algorithm` and `Best-response` Subroutines

Given Proposition 3.2, the remaining task is to instantiate the `Online-Algorithm` and `Best-response` subroutines. The requirements for `Online-Algorithm` are standard. Any no-regret

(6) online optimization algorithm is valid. Examples include Online Gradient Descent (Zinkevich, 2003), Exponentiated Gradient (Kivinen & Warmuth, 1997), and Follow-the-Regularized-Leader (Shalev-Shwartz, 2007). The `Best-response` subroutine, which corresponds to the dual function of the DRC-RL problem (3), is more difficult to instantiate and currently has no provable method for any specific uncertainty set among related works (Mankowitz et al., 2020; Wang et al., 2022; Bossens, 2023). Therefore, the key challenge is to efficiently and provably solve `Best-response` problem (5).

In detail, the `Best-response` problem with a given λ_t corresponds to the maximization problem of value functions of a form that often occurs in RL, e.g. $\max_\pi V_r^\pi$ in DR-RL, where

$$\pi_t \in \operatorname*{argmax}_{\pi\in\Pi} L(\pi,\lambda_t) = \operatorname*{argmax}_{\pi\in\Pi} V_r^\pi(\mu) - \lambda_t^\top V_g^\pi(\mu). \tag{7}$$

With a finite action space, similar maximization problems can be efficiently solved using iterative methods over greedy policies using some operators in various popular RL problems, such as standard RL (Scherrer et al., 2015), distributionally robust RL (Iyengar, 2005; Derman et al., 2021; Panaganti et al., 2022), constrained RL (Le et al., 2019; Miryoosefi et al., 2019), and regularized RL (Geist et al., 2019). Using a similar approach, for any policy π, we propose a consistency operator $\mathcal{T}^\pi : v \in \mathbb{R}^{\mathcal{S}} \mapsto \mathcal{T}^\pi v \in \mathbb{R}^{\mathcal{S}}$ so that for any given λ_t,

$$\begin{aligned}[\mathcal{T}^\pi v](s) = (r-\lambda_t^\top g)(s,\pi(s)) + \gamma\langle P^o_{s,\pi(s)}, v\rangle + \gamma \min_{P\in\mathcal{P}}\langle P_{s,\pi(s)} - P^o_{s,\pi(s)}, V_r^\pi\rangle \\ - \gamma\lambda_t^\top \min_{P\in\mathcal{P}}\langle P_{s,\pi(s)} - P^o_{s,\pi(s)}, V_g^\pi\rangle.\end{aligned} \tag{8}$$

where $(r-\lambda_t^\top g)(s,\pi(s)) \coloneqq r(s,\pi(s)) - \lambda_t^\top g(s,\pi(s))$ for the brevity sake.[2] Correspondingly, an optimality operator $\mathcal{T}^*$ with a fixed state s can be defined as $[\mathcal{T}^* v](s) = \max_{\pi\in\Pi}[\mathcal{T}^\pi v](s)$.

Proposition 3.3. *The consistency and optimality operators, i.e., $\mathcal{T}^\pi$ and $\mathcal{T}^*$, satisfy:*
(1) Monotonicity: let $v_1, v_2 \in \mathbb{R}^{\mathcal{S}}$ such that $v_1 \geq v_2$, then $\mathcal{T}^\pi v_1 \geq \mathcal{T}^\pi v_2$ and $\mathcal{T}^* v_1 \geq \mathcal{T}^* v_2$.
(2) Transition Invariance: for any $c \in \mathbb{R}$, we have $\mathcal{T}^\pi(v + c\mathbf{1}) = \mathcal{T}^\pi v + \gamma c\mathbf{1}$ and $\mathcal{T}^*(v + c\mathbf{1}) = \mathcal{T}^* v + \gamma c\mathbf{1}$.
(3) Contraction: The operator $\mathcal{T}^\pi$ and $\mathcal{T}^*$ are γ-contractions. Further, $V_r^\pi - \lambda_t^\top V_g^\pi$ is the unique stationary points of operator $\mathcal{T}^\pi$.

The properties summarized in Proposition 3.3 allow us to apply the consistency operator in an approximate modified policy iteration (AMPI) scheme (Scherrer et al., 2015) to solve `Best-response`. AMPI scheme generalizes both value iteration (as used in Section 2) and policy iteration methods and is widely used for other RL problems (Scherrer et al., 2015; Geist et al., 2019; Panaganti et al., 2022). The procedure of AMPI can be described as follows:

$$\pi^{k+1} = \arg\max_\pi {}^{\epsilon'_{k+1}}\mathcal{T}^\pi v^k \quad \text{and} \quad v^{k+1} = (\mathcal{T}^{\pi^{k+1}})^m v^k + \epsilon_{k+1}, \tag{9}$$

where $\epsilon_k \in \mathbb{R}^{\mathcal{S}}, \epsilon'_k \in \mathbb{R}^{\mathcal{S}}$ are some optimization errors in the k-th iteration. Here, we assume the operator $\max_\pi^{\epsilon'_{k+1}} \mathcal{T}^\pi v^k$ guarantees $\max_\pi[\mathcal{T}^\pi v^k](s) \leq [\mathcal{T}^{\pi^{k+1}} v^k](s) + \epsilon'_{k+1}(s)$ for all $s \in \mathcal{S}$ for now.

In words, the two update rules in (9) correspond to approximate policy improvement and approximate policy evaluation, respectively. In RL literature, those two steps can be solved by some oracles. Especially, the approximate policy improvement step is often represented as $\pi^{k+1} = \mathcal{G}^{\epsilon_{k+1}}(v^k)$, being the greedy policy with respect to v^k and an error term ϵ_{k+1} (Munos & Szepesvári, 2008; Lazaric et al., 2012; Scherrer et al., 2015; Geist et al., 2019). In these RL problems, such formulations, while being nominally different, coincide with ours in (9) as the greedy policy $\mathcal{T}^{\mathcal{G}(v^k)} v^k = \max_\pi \mathcal{T}^\pi v^k$ is optimal for the policy improvement step. Inspired by the literature, we first assume two oracles to execute these two steps for now, leading to the following assumption:

[2] Here V_r^π and V_g^π are the robust value functions that are fixed given π, making $\mathcal{T}^\pi$ in (8) not a practical operator yet until further specification.

Assumption 3.4. *There exist oracles that approximately solve (i) the policy improvement step with errors $\{\epsilon_k\}$, and (ii) the policy evaluation step in AMPI (9) with error $\{\epsilon'_k\}$.*

Assumption 3.4 also requires the existence of an ϵ'-approximated policy for policy improvement, which is not obviously valid as $\mathcal{T}^*$ may correspond to different best policies for different states. This issue is resolved if the greedy policy is optimal (as in Section 4) and is further discussed in Section 5. We also provide a standalone solution for this issue in Appendix D.1.

Now, under Assumption 3.4, for any k-th iteration, we are ready to control the loss $l_k := v^{\pi_t} - v^{\pi^k}$ via AMPI, where π_t is the solution to `Best-response` problem with respect to λ_t, v^π represents the unique stationary point of $\mathcal{T}^\pi$ for any policy π whose uniqueness is guaranteed by contraction in Proposition 3.3. The analysis is analogous to that in Scherrer et al. (2015).

Theorem 3.5. *Under Assumption 3.4, applying (9) for k-th iterations, the loss l_k satisfy,*

$$l_k \leq O(\gamma^k) + (2\bar{\epsilon}(\gamma - \gamma^k) + \bar{\epsilon}'(1-\gamma^k))/(1-\gamma)^2 \xrightarrow{k\to\infty} (2\bar{\epsilon}\gamma + \bar{\epsilon}')/(1-\gamma)^2, \tag{10}$$

where $\bar{\epsilon} \in \mathbb{R}^{\mathcal{S}}$ is the upperbound of errors $\{\epsilon_k\}$, i.e. $\forall k, \epsilon_k \leq \bar{\epsilon}$, and $\bar{\epsilon}' \in \mathbb{R}^{\mathcal{S}}$ is similarly defined as the upper bound of $\{\epsilon'_k\}$.

Theorem 3.5 shows that, when errors are relatively small, our AMPI (9) guarantees convergence to the solution of `Best-response` under Assumption 3.4. Combining this with a no-regret online algorithm, we complete the general framework for the DRC-RL problem as in Algorithm 1.

Finally, while the oracle for the approximate policy evaluation step can be implemented for several popular uncertainty sets (Shi et al., 2023; Clavier et al., 2023), the greedy policy solution for the approximate policy improvement step does not work, at least for our consistency operator $\mathcal{T}^\pi$. This is due to its dependency on the whole policy in its definition (8). Moreover, there is currently no provable efficient instantiation for general uncertainty set to enable the approximate policy improvement step in DRC-RL problems. Given the fact that `Best-response` corresponds to the fundamental dual function of (3), and that the policy improvement step consists of a popular class of iteration methods for `Best-response` type problems, we wonder:

(Q1) Can we design a specific uncertainty set for our operator $\mathcal{T}^\pi$ that enables solving DRC-RL without oracles, e.g. using greedy policies?

(Q2) Is it possible to design a better consistency operator that makes greedy policies optimal that in turn provably solves DRC-RL with our framework?

We address these two questions in the next two sections, respectively.

4 DRC-RL with R-Contamination Uncertainty Sets

In this section, we address (Q1) via a focus on the R-contamination uncertainty sets $\mathcal{P}_{s,a} := \{(1-\beta)P^o_{s,a} + \beta q \,|\, q \in \triangle(\mathcal{S})\}$ with a scalar robust level $\beta \in \mathbb{R}$. This uncertainty set has been studied in distributionally robust RL recently (Wang et al., 2022; Li & Lan, 2023). Considering this, we can simplify our consistency operator $\mathcal{T}^\pi$ without any loss in solving `Best-response` as

$$[\mathcal{T}^\pi v](s) = (r - \lambda_t^\top g)(s, \pi(s)) + \gamma(1-\beta)\langle P^o_{s,\pi(s)}, v\rangle + \gamma\beta(\min_{s'} V^\pi_r(s') - \lambda_t^\top \min_{s'} V^\pi_g(s')). \tag{11}$$

Please refer to Appendix B for detailed proof. Additionally, we adopt the following fail-state assumption (Panaganti et al., 2022).

Assumption 4.1 (fail-state)**.** *There is a fail state s_f for all the RMDPs, such that $r(s_f, a) = 0, g_i(s_f, a) = 0$ and $P_{s_f,a}(s_f) = 1$, for all $a \in \mathcal{A}$ and $P \in \mathcal{P}$.*

The fail-state assumption is commonly satisfied in practice as it corresponds to an end-game state in the simulator or real-world systems, in which all constraints are violated and the reward is zero.

Under this fail-state assumption, we always have $\min_{s'} V_r^\pi(s') = \min_{s'} V_g^\pi(s') = 0$, which makes the operator $\mathcal{T}^\pi$ correspond to a standard bellman consistency operator with shortened discount factor,

$$[\mathcal{T}^\pi v](s) = (r - \lambda_t^\top g)(s, \pi(s)) + \gamma(1-\beta)\langle P^0_{s,\pi(s)}, v\rangle. \tag{12}$$

Given that $\mathcal{T}^\pi$ takes the form of a standard consistency operator, the greedy policy is available for the policy improvement step. Thus, any instantiation of AMPI (9), such as simple value iteration or policy iteration, can efficiently and provably solve `Best-response` problem under small errors. The distributionally robust constrained problem therefore has a provable solution in the case of the R-contamination uncertainty set. The details of an instantiation of `Best-response` and `Online-Algorithm` are discussed in Appendix D.2.

Our result for the R-contamination uncertainty set indicates that a smaller discount factor, i.e. a smaller effective horizon, gives higher distributional robustness. However, as this smaller discount factor is a consequence of the robustness objective, the discount factor should remain unchanged, when designing constraints threshold τ and testing in shifted environments. Such an unchanged discount factor in thresholds and tests differs our solution from simply scaling the problem.

Finally, it is worth noting that the case of no constraints, e.g. $g = \tau = 0$, implies that our analysis also gives a provable solution to the distributionally robust RL problem on R-contamination uncertainty sets.

5 On the Intractability of Greedy Policies for DRC-RL

In this section, we answer question (Q2) with a negative result showing that the combination of constraints and distributional robustness requires different algorithmic tools than either robust RL or constrained RL do alone. In detail, we show that for any 'good' consistency operator, the greedy policies are not generally optimal for the policy improvement step in (9). This, in turn, prevents any algorithm from a popular class of iteration methods from tractably solving the DRC-RL problem.

To begin, we formally define the optimality of greedy policy and connect it to the operator. We assume policy $\pi \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ includes both deterministic and stochastic policies.

Definition 5.1 (Greedy Policy Enabling)**.** *We state an consistency operator $\mathcal{T}^\pi$ enables the greedy policy if there exist a function $g : \mathbb{R}^{\mathcal{S}} \times \mathcal{A} \to \mathbb{R}^{\mathcal{S}}$ such that $\forall v \in \mathbb{R}^{\mathcal{S}}, s \in \mathcal{S}, \max_{a\in\mathcal{A}} g(v,a)[s] = \max_{\pi\in\Pi}[\mathcal{T}^\pi v](s)$, i.e. greedy policies are optimal.*

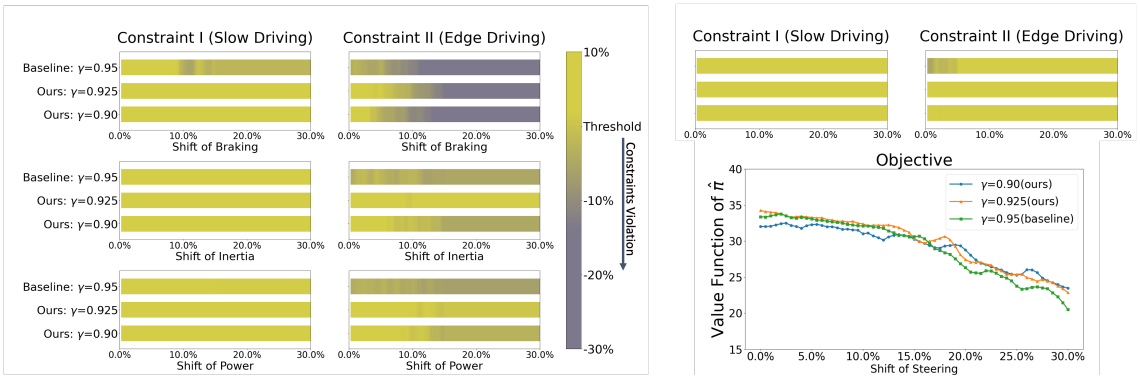
Definition 5.2 (Operator Linearity)**.** *The consistency operator $\mathcal{T}^\pi$, that takes policy π as an input, is linear if there exists a function $f : \mathbb{R}^{\mathcal{S}} \to \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ that is independent of π, such that $\forall v \in \mathbb{R}^{\mathcal{S}}, s \in \mathcal{S}$, we have $[\mathcal{T}^\pi v](s) = \langle \pi[s,\cdot], f(v)[s,\cdot]\rangle = \langle \pi, f(v)\rangle_s$.*

Take distributional robust RL as an example, its robust consistency operator (2): $[\mathcal{T}^\pi_{r,\text{rob}} v](s) = \langle \pi[s,\cdot], f(v)[s,\cdot]\rangle$ is linear where $f(v)[s,a] = r_{s,a} + \min_{P\in\mathcal{P}}\langle P_{s,a}, v\rangle$. Thus, one only needs to treat $f(v)[s,a]$ as $g(v,a)[s]$ in Definition 5.1 to enable the greedy policy. It is not hard to find that a linear operator naturally enables the greedy policy as in the above example, and we generalize this in the following result. All proofs for this section are presented in Appendix C.

Lemma 5.3. *The linear operator is equivalent to greedy policy-enabled operators in Definition 5.1 in the following ways: (i) If a consistency operator is a linear operator, then it enables the greedy policy. (ii) If a consistency operator $\mathcal{T}^\pi$ enables the greedy policy, then there always exists a linear operator $\mathcal{T}^\pi_{linear}$ that can substitute $\mathcal{T}^\pi$ without any loss in policy improvement step, i.e. $\forall v \in \mathbb{R}^{\mathcal{S}}, s \in \mathcal{S}, \max_{\pi\in\Pi}[\mathcal{T}^\pi v](s) = \max_{\pi\in\Pi}[\mathcal{T}^\pi_{linear} v](s)$.*

However, the following shows that it is impossible to have a consistency operator that simultaneously converges as a contraction to our target in `Best-response` and enables the greedy policy.

Theorem 5.4. *There is no consistency operator $\mathcal{T}^\pi$, that takes any policy π as an input, simultaneously satisfies for given γ and every λ_t: (i) Linearity. (ii) γ-Contraction to our target: the operator $\mathcal{T}^\pi$ is a contraction such that for every policy $\pi \in \Pi$, $V_r^\pi - \lambda_t^\top V_g^\pi$ is the unique stationary point of $\mathcal{T}^\pi$.*



(a) Constraints Satisfaction under Shifts

(b) Performance under Steering Shift

Figure 1: The four bar graphs denote the constraints satisfaction (green means satisfied) when shifts of power, inertia, braking magnitude, and steering angle occur. The lower right figure indicates the value of the objective (higher is better) when the steering angle is shifted. All evaluations are based on the value function (accumulated rewards) of mixture policy $\hat{\pi}$.

Corollary 5.5. *There is no consistency operator $\mathcal{T}^\pi$ that enables greedy policy while retaining as a γ-contraction to $V_r^\pi - \lambda_t^\top V_g^\pi$.*

Theorem 5.4 and Corollary 5.5 highlight the additional difficulty of DRC-RL as compared to robust or constrained RL, where iterative methods can be successful. Our proof (Appendix C.2) demonstrates how this difficulty arises from the combination of constraints and distributional robustness.

Comparing our previous success in the case of R-contamination sets (Section 4) to these impossible results, it becomes evident that the additional fail-state assumption is critical. This assumption restricts the possible value function space and provides an additional transition structure that avoids the challenges underlying Theorem 5.4.

Following the fail-state assumption, we believe it is possible to design additional conditions for other forms of uncertainty sets that resolve such impossibilities. Although we have not yet found clear and rigorous guidance, we tentatively acknowledge that the absence of linearity in worst-case transition kernels is essential for Theorem 5.4. Therefore a structured value function space or an augmented state space (such as in Sootla et al. (2022)) might be helpful to give tractable solutions with our framework in Section 3.

6 Experiments and Evaluation

In this section, we present a focused experiment to validate our solution in Section 4. This solution specializes to the case of constrained RL in Le et al. (2019) when setting the robustness level to 0. We present critical settings here and refer to Appendix E for more details.

Task Setting. We choose the high-dimensional Car Racing task (Towers et al., 2023) where the agents must traverse as far as possible on track, with each reward for a passed tile of track and a small negative reward for each second. Two constraints are designed: slow driving and edge driving. States off the track excessively are considered fail-states.

Algorithms Design and Baseline. We adopt Fitted Q Iteration (Ernst et al., 2005) and Exponentiated Gradient (Kivinen & Warmuth, 1997) for `Best-response` and `Online-Algorithm`. Evaluations are executed using the simulator and meet the requirements of Proposition 3.2. We choose $\gamma = 0.95$ as the initial discount factor for the baseline, since in this case our solution retrieves the constrained RL algorithm in Le et al. (2019) and is considered as zero robustness level. We select two robustness levels β for evaluation. These result in discount factors $\gamma(1-\beta) = \{0.90, 0.925\}$,

and correspond to a maximized 50% decrease of the effective horizon. All other hyperparameters, including random seeds, do not change across γ.

Procedure and Criteria. For each γ, 25 rounds are executed, which guarantees a duality gap of less than .01. The mixture policy $\hat{\pi} = \sum_{t=1}^{25} \frac{1}{25}\pi_t$ is then tested in the shifted environments. When testing with a shifted environment, the objective and two constraints for each mixture policy are measured with value function and initial discount factor $\gamma = 0.95$. All results in shifted environments are smoothed as the mean of 3 random seeds and a 1% shift window. We regard constraints as hard and prioritize constraint satisfaction as the main criterion for robustness, with the value of the objective as the secondary criterion.

Results. Our experiments verify that smaller γ learns a more robust policy in the car-racing example. Results are shown in Figure 1: In the left plot, we present constraint satisfaction under shifts of power, inertia, and brake magnitude. Our learned policies with smaller discount factors perform better. In the right plot, we present a full set of evaluations when the steering angle is shifted. Here the learned policies with smaller discount factors not only satisfy more constraints but may also have better objective value when a certain shift occurs.

7 Related Works

Constrained RL. Constrained RL aims at maximizing expected cumulative reward while adhering to specified constraints. Applications of Constrained RL cover a wide array of topics, such as resource allocation for numerous users in grid systems (Wang et al., 2020; De Nijs et al., 2021; Mo et al., 2023), human satisfaction in human-robot interaction (El-Shamouty et al., 2020; Liu et al., 2023), and the safety level of robotic agents (Wachi & Sui, 2020; Zhang et al., 2020; Brunke et al., 2022; Tambon et al., 2022). The underlying decision-making problem in constrained RL can be represented as a constrained Markov decision process (Altman, 2021) that has a bilevel structure with strong duality (Miryoosefi et al., 2019; Paternain et al., 2019). Most works in constrained RL employ model-based methods (Efroni et al., 2020; Bura et al., 2022). To develop model-free and policy gradient methods, many additional constrained RL algorithms, involving online (Ding et al., 2021; Wachi et al., 2021) or offline (Le et al., 2019) interactions, embrace a primal-dual methodology.

Distributionally Robust RL. Distributionally robust RL tackles the challenge of formulating a policy resilient to shifts between training and testing environments by using robust Markov decision process (Nilim & Ghaoui, 2003; Iyengar, 2005) as the underlying decision-making problem. Prior works (Xu & Mannor, 2010; Wiesemann et al., 2013; Yu & Xu, 2015; Mannor et al., 2016; Russel & Petrik, 2019) have shown that distributional robustness is essential when the environment shifts. Recently, work has also started to provide a concrete theoretical understanding of distributionally robust RL about the convergence of algorithms and sample complexity (Yang et al., 2021b; Panaganti & Kalathil, 2021b; Zhou et al., 2021; Shi & Chi, 2022; Wang et al., 2023a; Blanchet et al., 2023; Liu et al., 2022; Wang et al., 2023c; Liang et al., 2023; Shi et al., 2023; Wang & Zou, 2021; Xu et al., 2023; Dong et al., 2022; Ramesh et al., 2023; Panaganti et al., 2022; Ma et al., 2022; Wang et al., 2023b; Li et al., 2022; Kumar et al., 2023; Clavier et al., 2023; Yang et al., 2023; Zhang et al., 2023; Li & Lan, 2023; Wang et al., 2024; Yang et al., 2021a; Panaganti & Kalathil, 2022; Shi et al., 2024). Additionally, other recent work (Panaganti & Kalathil, 2021a; Wang & Zou, 2021; Panaganti et al., 2022; 2024) has employed general function approximation to devise model-free online and offline robust RL algorithms.

Related work on DRC-RL. The study of DRC-RL can be traced back to at least Russel et al. (2020) and Mankowitz et al. (2020), where the basic formulation is proposed and first-order methods, such as Robust Constrained Policy Gradient (RCPG), are studied. More recently, Wang et al. (2022) focuses on the dual problem and proposes a first-order method that achieves the convergence guarantee to a stationary point with additional approximation, and Bossens (2023) extends RCPG to Lagrangian or adversarial updates for the dual problem of a different formulation. A simultaneous work of ours is Sun et al. (2024), where they proposed a projected gradient descend style algorithm that guarantees per-step improvement and constraints violation. However, these existing works do

not provide provable guarantees for an end-to-end framework of DRC-RL, which instead makes relaxations for devising tractable objectives.

Beyond the distributional robustness we studied, a variety of relaxations quantifying robustness, such as assuming a stochastic setting with uncertainty on the distribution of kernels (Queeney & Benosman, 2024) or considering risk-averse constraints (Kim et al., 2024), are studied. Such relaxation provides a more informative structure to avoid the minimax formulation, and is therefore beyond the scope of this paper.

8 Limitations and Future Directions

This paper is initiated from the primal-dual framework upon mixed policies, and derives algorithmic and theoretical results that hold independently for every DRC-RL problem. While we highlight these results mainly presented in Section 4 and Section 5, we consider our framework immature and has certain limitations. In specific, the use of mixed policies may be inefficient and has a large variance under resource limitation. Any single policy inside the resulting mixed policies, on the other hand, might not satisfy the constraints when executed consistently. Moreover, the intractability of greedy policies indicates that the primal-dual paradigm might not fit with DRC-RL problems.

As per, we regard applying other methods, such as epi-graph methods, as a future direction to fully solve DRC-RL problems. Another interesting direction is to study additional structures, as provided by the fail-state assumption, to encompass both distributional robustness and constraint satisfaction upon the proposed framework. Because of the fundamental role of iterative methods in RL, such structures should be significant despite the context and may correspond to other findings in related fields such as neuroscience.

9 Conclusion

In this paper, we present a primal-dual algorithmic framework for the distributionally robust constrained RL problem (DRC-RL). Our framework provides the first efficient provable solution for R-contamination uncertainty sets. We additionally prove the intractability of greedy policies for general uncertainty sets, which prevents the use of popular iterative methods unless the uncertainty sets and additional assumptions maintain additional structures. In the case of R-contamination uncertainty sets, a simple rule relating the discount factor and distributional robustness is discovered. We view this paper as a stepping stone to explore the theoretical and algorithmic understanding of distributional robustness and constraints in RL, and anticipate future research stemming from our framework and results.

10 Acknowledgment

The work is supported by the NSF through CNS-2146814, CPS-2136197, CNS-2106403, NGSDI-2105648, CCF-1918865, and by the Resnick Sustainability Institute at Caltech, and by the gift from Latitude AI. Z. Zhang acknowledges support from the Tsien's Excellence in Engineering Program at Tsinghua. K. Panaganti acknowledges support from the 'PIMCO Postdoctoral Fellow in Data Science' fellowship at Caltech. The work of L. Shi is supported in part by the Resnick Institute and Computing, Data, and Society Postdoctoral Fellowship at Caltech. The authors thank anonymous RLC 2024 reviewers for their constructive comments on an earlier draft of this paper.

References

Eitan Altman. *Constrained Markov decision processes*. Routledge, 2021.

Jessamy Bagenal, Marco Zenone, Nason Maani, and Skye Barbic. Embracing the non-traditional: alcohol advertising on tiktok. *BMJ Global Health*, 8(1), 2023.

Jose Blanchet, Miao Lu, Tong Zhang, and Han Zhong. Double pessimism is provably efficient for distributionally robust offline reinforcement learning: Generic algorithm and robust partial coverage. *arXiv preprint arXiv:2305.09659*, 2023.

David M Bossens. Robust lagrangian and adversarial policy gradient for robust constrained markov decision processes. *arXiv preprint arXiv:2308.11267*, 2023.

Stephen P Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Lukas Brunke, Melissa Greeff, Adam W Hall, Zhaocong Yuan, Siqi Zhou, Jacopo Panerati, and Angela P Schoellig. Safe learning in robotics: From learning-based control to safe reinforcement learning. *Annual Review of Control, Robotics, and Autonomous Systems*, 5:411–444, 2022.

Archana Bura, Aria HasanzadeZonuzy, Dileep Kalathil, Srinivas Shakkottai, and Jean-Francois Chamberland. Dope: Doubly optimistic and pessimistic exploration for safe reinforcement learning. *Advances in Neural Information Processing Systems*, 35:1047–1059, 2022.

Jie Chen, Ron J Patton, and Hong-Yue Zhang. Design of unknown input observers and robust fault detection filters. *International Journal of control*, 63(1):85–105, 1996.

Pierre Clavier, Erwan Le Pennec, and Matthieu Geist. Towards minimax optimality of model-based robust reinforcement learning. *arXiv preprint arXiv:2302.05372*, 2023.

Frits De Nijs, Erwin Walraven, Mathijs De Weerdt, and Matthijs Spaan. Constrained multiagent markov decision processes: A taxonomy of problems and algorithms. *Journal of Artificial Intelligence Research*, 70:955–1001, 2021.

Esther Derman, Matthieu Geist, and Shie Mannor. Twice regularized mdps and the equivalence between robustness and regularization. *Advances in Neural Information Processing Systems*, 34: 22274–22287, 2021.

Dongsheng Ding, Xiaohan Wei, Zhuoran Yang, Zhaoran Wang, and Mihailo Jovanovic. Provably efficient safe exploration via primal-dual policy optimization. In *International Conference on Artificial Intelligence and Statistics*, pp. 3304–3312. PMLR, 2021.

Jing Dong, Jingwei Li, Baoxiang Wang, and Jingzhao Zhang. Online policy optimization for robust MDP. *arXiv preprint arXiv:2209.13841*, 2022.

Yonathan Efroni, Shie Mannor, and Matteo Pirotta. Exploration-exploitation in constrained mdps. *arXiv preprint arXiv:2003.02189*, 2020.

Mohamed El-Shamouty, Xinyang Wu, Shanqi Yang, Marcel Albus, and Marco F Huber. Towards safe human-robot collaboration using deep reinforcement learning. In *2020 IEEE international conference on robotics and automation (ICRA)*, pp. 4899–4905. IEEE, 2020.

Damien Ernst, Pierre Geurts, Louis Wehenkel, and Michael L. Littman. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6(4):503 – 556, 2005. ISSN 15324435.

Yoav Freund and Robert E Schapire. Adaptive game playing using multiplicative weights. *Games and Economic Behavior*, 29(1-2):79–103, 1999.

Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. A theory of regularized markov decision processes. In *International Conference on Machine Learning*, pp. 2160–2169. PMLR, 2019.

Sami Haddadin, Simon Haddadin, Augusto Khoury, Tim Rokahr, Sven Parusel, Rainer Burgkart, Antonio Bicchi, and Alin Albu-Schäffer. A truly safely moving robot has to know what injury it may cause. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5406–5413. IEEE, 2012.

Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

Peter J Huber. A robust version of the probability ratio test. *The Annals of Mathematical Statistics*, pp. 1753–1758, 1965.

Garud N Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2): 257–280, 2005.

Dohyeong Kim, Kyungjae Lee, and Songhwai Oh. Trust region-based safe distributional reinforcement learning for multiple constraints. *Advances in neural information processing systems*, 36, 2024.

Jyrki Kivinen and Manfred K Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *information and computation*, 132(1):1–63, 1997.

MSW KRM, Hannah Kia, and Ashley Lacombe-Duncan. Examining tiktok's potential for community-engaged digital knowledge mobilization with equity-seeking groups. *J Med Internet Res*, 23(12):e30315, 2021.

Navdeep Kumar, Esther Derman, Matthieu Geist, Kfir Levy, and Shie Mannor. Policy gradient for s-rectangular robust markov decision processes. *arXiv preprint arXiv:2301.13589*, 2023.

Alessandro Lazaric, Mohammad Ghavamzadeh, and Rémi Munos. Finite-sample analysis of least-squares policy iteration. *Journal of Machine Learning Research*, 13(Oct):3041–3074, 2012.

Hoang Le, Cameron Voloshin, and Yisong Yue. Batch policy learning under constraints. In *International Conference on Machine Learning*, pp. 3703–3712. PMLR, 2019.

Yan Li and Guanghui Lan. First-order policy optimization for robust policy evaluation. *arXiv preprint arXiv:2307.15890*, 2023.

Yan Li, Guanghui Lan, and Tuo Zhao. First-order policy optimization for robust markov decision process. *arXiv preprint arXiv:2209.10579*, 2022.

Zhipeng Liang, Xiaoteng Ma, Jose Blanchet, Jiheng Zhang, and Zhengyuan Zhou. Single-trajectory distributionally robust reinforcement learning. *arXiv preprint arXiv:2301.11721*, 2023.

Puze Liu, Kuo Zhang, Davide Tateo, Snehal Jauhri, Zhiyuan Hu, Jan Peters, and Georgia Chalvatzaki. Safe reinforcement learning of dynamic high-dimensional robotic tasks: navigation, manipulation, interaction. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9449–9456. IEEE, 2023.

Zijian Liu, Qinxun Bai, Jose Blanchet, Perry Dong, Wei Xu, Zhengqing Zhou, and Zhengyuan Zhou. Distributionally robust q-learning. In *International Conference on Machine Learning*, pp. 13623–13643. PMLR, 2022.

Xiaoteng Ma, Zhipeng Liang, Jose Blanchet, Mingwen Liu, Li Xia, Jiheng Zhang, Qianchuan Zhao, and Zhengyuan Zhou. Distributionally robust offline reinforcement learning with linear function approximation. *arXiv preprint arXiv:2209.06620*, 2022.

Daniel J. Mankowitz, Nir Levine, Rae Jeong, Abbas Abdolmaleki, Jost Tobias Springenberg, Yuanyuan Shi, Jackie Kay, Todd Hester, Timothy Mann, and Martin Riedmiller. Robust reinforcement learning for continuous control with model misspecification. In *International Conference on Learning Representations*, 2020.

Shie Mannor, Ofir Mebel, and Huan Xu. Robust mdps with k-rectangular uncertainty. *Mathematics of Operations Research*, 41(4):1484–1509, 2016.

Douglas Maraun. Bias correcting climate change simulations-a critical review. *Current Climate Change Reports*, 2:211–220, 2016.

Sobhan Miryoosefi, Kianté Brantley, Hal Daume III, Miro Dudik, and Robert E Schapire. Reinforcement learning with convex constraints. *Advances in neural information processing systems*, 32, 2019.

Fan Mo, Hamood Ur Rehman, Fabio Marco Monetti, Jack C Chaplin, David Sanderson, Atanas Popov, Antonio Maffei, and Svetan Ratchev. A framework for manufacturing system reconfiguration and optimisation utilising digital twins and modular artificial intelligence. *Robotics and Computer-Integrated Manufacturing*, 82:102524, 2023.

Rémi Munos. Performance bounds in l_p-norm for approximate value iteration. *SIAM journal on control and optimization*, 46(2):541–561, 2007.

Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(5), 2008.

Arnab Nilim and Laurent Ghaoui. Robustness in markov decision problems with uncertain transition matrices. *Advances in neural information processing systems*, 16, 2003.

Kishan Panaganti and Dileep Kalathil. Robust reinforcement learning using least squares policy iteration with provable performance guarantees. In *International Conference on Machine Learning (ICML)*, pp. 511–520, 2021a.

Kishan Panaganti and Dileep Kalathil. Sample complexity of model-based robust reinforcement learning. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pp. 2240–2245. IEEE, 2021b.

Kishan Panaganti and Dileep Kalathil. Sample complexity of robust reinforcement learning with a generative model. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, 2022. URL https://proceedings.mlr.press/v151/panaganti22a.html.

Kishan Panaganti, Zaiyan Xu, Dileep Kalathil, and Mohammad Ghavamzadeh. Robust reinforcement learning using offline data. *Advances in neural information processing systems*, 35:32211–32224, 2022.

Kishan Panaganti, Adam Wierman, and Eric Mazumdar. Model-free robust ϕ-divergence reinforcement learning using both offline and online data. In *International Conference on Machine Learning*. PMLR, 2024.

Santiago Paternain, Luiz Chamon, Miguel Calvo-Fullana, and Alejandro Ribeiro. Constrained reinforcement learning has zero duality gap. *Advances in Neural Information Processing Systems*, 32, 2019.

Nicholas J Pioch, James Melhuish, Andy Seidel, Eugene Santos Jr, Deqing Li, and Mark Gorniak. Adversarial intent modeling using embedded simulation and temporal bayesian knowledge bases. In *Modeling and Simulation for Military Operations IV*, volume 7348, pp. 115–126. SPIE, 2009.

James Queeney and Mouhacine Benosman. Risk-averse model uncertainty for distributionally robust safe reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2024.

Shyam Sundhar Ramesh, Pier Giuseppe Sessa, Yifan Hu, Andreas Krause, and Ilija Bogunovic. Distributionally robust model-based reinforcement learning with large state spaces. *arXiv preprint arXiv:2309.02236*, 2023.

Reazul Hasan Russel and Marek Petrik. Beyond confidence regions: Tight bayesian ambiguity sets for robust mdps. *Advances in Neural Information Processing Systems*, 2019.

Reazul Hasan Russel, Mouhacine Benosman, and Jeroen Van Baar. Robust constrained-mdps: Soft-constrained robust policy optimization under model uncertainty. *arXiv preprint arXiv:2010.04870*, 2020.

Bruno Scherrer, Mohammad Ghavamzadeh, Victor Gabillon, Boris Lesner, and Matthieu Geist. Approximate modified policy iteration and its application to the game of tetris. *J. Mach. Learn. Res.*, 16(49):1629–1676, 2015.

Shai Shalev-Shwartz. *Online learning: Theory, algorithms, and applications*. Hebrew University, 2007.

Laixi Shi and Yuejie Chi. Distributionally robust model-based offline reinforcement learning with near-optimal sample complexity. *arXiv preprint arXiv:2208.05767*, 2022.

Laixi Shi, Gen Li, Yuting Wei, Yuxin Chen, Matthieu Geist, and Yuejie Chi. The curious price of distributional robustness in reinforcement learning with a generative model. *arXiv preprint arXiv:2305.16589*, 2023.

Laixi Shi, Eric Mazumdar, Yuejie Chi, and Adam Wierman. Sample-efficient robust multi-agent reinforcement learning in the face of environmental uncertainty. *arXiv preprint arXiv:2404.18909*, 2024.

Aivar Sootla, Alexander I Cowen-Rivers, Taher Jafferjee, Ziyan Wang, David H Mguni, Jun Wang, and Haitham Ammar. Sauté rl: Almost surely safe reinforcement learning using state augmentation. In *International Conference on Machine Learning*, pp. 20423–20443. PMLR, 2022.

Zhongchang Sun, Sihong He, Fei Miao, and Shaofeng Zou. Constrained reinforcement learning under model mismatch. *arXiv preprint arXiv:2405.01327*, 2024.

Florian Tambon, Gabriel Laberge, Le An, Amin Nikanjam, Paulina Stevia Nouwou Mindom, Yann Pequignot, Foutse Khomh, Giulio Antoniol, Ettore Merlo, and François Laviolette. How to certify machine learning based safety-critical systems? a systematic literature review. *Automated Software Engineering*, 29(2):38, 2022.

Mark Towers, Jordan K. Terry, Ariel Kwiatkowski, John U. Balis, Gianluca de Cola, Tristan Deleu, Manuel Goulão, Andreas Kallinteris, Arjun KG, Markus Krimmel, Rodrigo Perez-Vicente, Andrea Pierré, Sander Schulhoff, Jun Jet Tai, Andrew Tan Jin Shen, and Omar G. Younis. Gymnasium, March 2023. URL https://zenodo.org/record/8127025.

Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, 2019.

Akifumi Wachi and Yanan Sui. Safe reinforcement learning in constrained markov decision processes. In *International Conference on Machine Learning*, pp. 9797–9806. PMLR, 2020.

Akifumi Wachi, Yunyue Wei, and Yanan Sui. Safe policy optimization with local generalized linear function approximations. *Advances in Neural Information Processing Systems*, 34:20759–20771, 2021.

He Wang, Laixi Shi, and Yuejie Chi. Sample complexity of offline distributionally robust linear markov decision processes. *arXiv preprint arXiv:2403.12946*, 2024.

Shengbo Wang, Nian Si, Jose Blanchet, and Zhengyuan Zhou. A finite sample complexity bound for distributionally robust Q-learning. *arXiv preprint arXiv:2302.13203*, 2023a.

Shengbo Wang, Nian Si, Jose Blanchet, and Zhengyuan Zhou. On the foundation of distributionally robust reinforcement learning. *arXiv preprint arXiv:2311.09018*, 2023b.

Shengbo Wang, Nian Si, Jose Blanchet, and Zhengyuan Zhou. Sample complexity of variance-reduced distributionally robust Q-learning. *arXiv preprint arXiv:2305.18420*, 2023c.

Wei Wang, Nanpeng Yu, Yuanqi Gao, and Jie Shi. Safe off-policy deep reinforcement learning algorithm for volt-var control in power distribution systems. *IEEE Transactions on Smart Grid*, 11(4):3008–3018, 2020. doi: 10.1109/TSG.2019.2962625.

Yue Wang and Shaofeng Zou. Online robust reinforcement learning with model uncertainty. *Advances in Neural Information Processing Systems*, 34:7193–7206, 2021.

Yue Wang and Shaofeng Zou. Policy gradient method for robust reinforcement learning. In *International Conference on Machine Learning*, pp. 23484–23526. PMLR, 2022.

Yue Wang, Fei Miao, and Shaofeng Zou. Robust constrained reinforcement learning. *arXiv preprint arXiv:2209.06866*, 2022.

Carlo Weidemann, Nils Mandischer, Frederick van Kerkom, Burkhard Corves, Mathias Hüsing, Thomas Kraus, and Cyryl Garus. Literature review on recent trends and perspectives of collaborative robotics in work 4.0. *Robotics*, 12(3):84, 2023.

Wolfram Wiesemann, Daniel Kuhn, and Berç Rustem. Robust Markov decision processes. *Mathematics of Operations Research*, 38(1):153–183, 2013.

Huan Xu and Shie Mannor. Distributionally robust Markov decision processes. In *Advances in Neural Information Processing Systems*, pp. 2505–2513, 2010.

Zaiyan Xu, Kishan Panaganti, and Dileep Kalathil. Improved sample complexity bounds for distributionally robust reinforcement learning. *arXiv preprint arXiv:2303.02783*, 2023.

Wenhao Yang, Liangyu Zhang, and Zhihua Zhang. Towards theoretical understandings of robust markov decision processes: Sample complexity and asymptotics, 2021a. URL https://arxiv.org/abs/2105.03863.

Wenhao Yang, Liangyu Zhang, and Zhihua Zhang. Towards theoretical understandings of robust markov decision processes: Sample complexity and asymptotics. *arXiv preprint arXiv:2105.03863*, 2021b.

Wenhao Yang, Han Wang, Tadashi Kozuno, Scott M Jordan, and Zhihua Zhang. Avoiding model estimation in robust markov decision processes with a generative model. *arXiv preprint arXiv:2302.01248*, 2023.

Pengqian Yu and Huan Xu. Distributionally robust counterpart in Markov decision processes. *IEEE Transactions on Automatic Control*, 61(9):2538–2543, 2015.

Jesse Zhang, Brian Cheung, Chelsea Finn, Sergey Levine, and Dinesh Jayaraman. Cautious adaptation for reinforcement learning in safety-critical settings. In *International Conference on Machine Learning*, pp. 11055–11065. PMLR, 2020.

Runyu Zhang, Yang Hu, and Na Li. Regularized robust mdps and risk-sensitive mdps: Equivalence, policy gradient, and sample complexity. *arXiv preprint arXiv:2306.11626*, 2023.

Zhengqing Zhou, Qinxun Bai, Zhengyuan Zhou, Linhai Qiu, Jose Blanchet, and Peter Glynn. Finite-sample regret bound for distributionally robust offline tabular reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 3331–3339, 2021.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th international conference on machine learning (icml-03)*, pp. 928–936, 2003.

A Proof for Section 3: DRC-RL with General Uncertainty Sets

A.1 Proof of Proposition 3.1

We assume $\pi_{\alpha,\{\pi_i\}_{i=1}^T}, \pi'_{\alpha',\{\pi_i\}_{i=1}^T} \in Conv(\Pi)$ are two mixed policies defined in Section 3.1 with the same candidate deterministic policy set without loss of generality (Otherwise we could simply combine two sets and set zero to new candidate policy for each categorical distribution). For brevity, we use π_α and $\pi_{\alpha'}$ to denote them.

First, by the definition of mixed policy in Section 3.1, $Conv(\Pi)$ is indeed a convex hull that linear combination of policies over $c \in [0, 1]$ satisfies,

$$\pi_{c\alpha+(1-c)\alpha'} = c\pi_\alpha + (1-c)\pi_{\alpha'} \in Conv(\Pi). \tag{13}$$

Then we show that the robust value function is linear to mixed policy, with $c \in [0, 1]$,

$$\begin{aligned} V_r^{\pi_{c\alpha+(1-c)\alpha'}} &= \sum_\pi (c\alpha(\pi) + (1-c)\alpha'(\pi))V_r^\pi && (14)\\ &= c\sum_\pi \alpha(\pi)V_r^\pi + (1-c)\sum_\pi \alpha'(\pi)V_r^\pi && (15)\\ &= cV^{\pi_\alpha} + (1-c)V^{\pi_{\alpha'}}. && (16)\end{aligned}$$

where the first equality comes from the definition of the robust value function of a mixed policy in Section 3.1.

Naturally, the robust value function for rewards of constraints V_g^π is also linear to mixed policy.

Therefore, the constrained optimization problem (3) becomes convex, and hence the strong duality holds with Slater's condition (Boyd & Vandenberghe, 2004).

A.2 Proof of Proposition 3.2

First, as proved in A.1, Lagrangian $L(\pi, \lambda) = V_r^\pi - \lambda^T V_g^\pi$ is linear to both policy π and multiplier λ. (We are treating Π as its convex hull $Conv(\Pi)$ now.)

When `Online-Algorithm` is chosen as a no-regret online learning algorithm with the negative Lagrangian $-L(\pi, \lambda)$ as loss (Kivinen & Warmuth, 1997; Zinkevich, 2003), we have

$$\sum_t (-L)(\pi_t, \lambda_t) \geq \max_\lambda \sum_t (-L)(\pi_t, \lambda) - o(T) \tag{17}$$

Then, recalling that π_t is the `Best-response` (5) given current λ_t, we have

$$\begin{aligned} \min_\lambda L(\hat{\pi}_T, \lambda) &= \min_\lambda \frac{1}{T}\sum_t L(\pi_t, \lambda) && (18)\\ &= -\max_\lambda \frac{1}{T}\sum_t -L(\pi_t, \lambda) && (19)\\ &\overset{\text{(i)}}{\geq} \frac{1}{T}\sum_t L(\pi_t, \lambda_t) - \frac{o(T)}{T} && (20)\\ &\overset{\text{(ii)}}{\geq} \frac{1}{T}\sum_t L(\pi, \lambda_t) - \frac{o(T)}{T} \quad , \forall \pi \in \Pi && (21)\\ &= L(\pi, \hat{\lambda}_T) - \frac{o(T)}{T} \quad , \forall \pi \in \Pi && (22)\end{aligned}$$

where the first and the last equalities come from the linearity of the Lagrangian w.r.t. the policy and the Lagrange multiplier, and the definition of $\hat{\pi}_T$ and $\hat{\lambda}_T$ in meta algorithm 1; (i) holds by Eq.(17), and (ii) arises from the fact that π_t is the `Best-response` (5)

Additionally, inserting the fact $\max_\pi L(\pi, \hat{\lambda}_T) \geq L(\hat{\pi}_T, \hat{\lambda}_T) \geq \min_\lambda L(\hat{\pi}_T, \lambda)$ one has

$$\max_\pi L(\pi, \hat{\lambda}_T) \geq \min_\lambda L(\hat{\pi}_T, \lambda) \geq \max_\pi L(\pi, \hat{\lambda}_T) - \frac{o(T)}{T}. \tag{23}$$

Finally, recalling that $L_{\max} = \max_\pi L(\pi, \hat{\lambda}_T)$ and $L_{\min} = \min_\lambda L(\hat{\pi}_T, \lambda)$ in meta algorithm 1. The duality gap $L_{\max} - L_{\min}$ is bound to smaller than positive threshold ω. The exact convergence rate of Algorithm 1 will depend on the choice of `Online-Algorithm`. For example, the algorithm will terminate after $\mathcal{O}(\frac{1}{\omega^2})$ rounds if online algorithms with regret scaling as $\Omega(\sqrt{T})$ are chosen (such as online gradient descent with regularizer). □

A.3 Proof for Proposition 3.3

Through out this proof, we denote P^o_π as the vector $[P^o_{s,\pi(s)}]^\top_{s\in\mathcal{S}}$, and $v_1, v_2 \in \mathbb{R}^{\mathcal{S}}, s \in \mathcal{S}$.

Monotonicity. By the definition of $\mathcal{T}^\pi$, we have with $v_1 \geq v_2$,

$$\mathcal{T}^\pi v_1 - \mathcal{T}^\pi v_2 = \gamma\langle P^o_\pi, v_1 - v_2\rangle \geq 0. \tag{24}$$

Then for $\mathcal{T}^*$, we denote $\pi^s_1 \coloneqq \operatorname{argmax}_{\pi\in\Pi} \mathcal{T}^\pi v_1(s)$ and similar for π^s_2, we then have for every state s,

$$\mathcal{T}^* v_1(s) - \mathcal{T}^* v_2(s) = \mathcal{T}_s^{\pi^s_1} v_1(s) - \mathcal{T}_s^{\pi^s_2} v_2(s) \geq \mathcal{T}_s^{\pi^s_1} v_1(s) - \mathcal{T}_s^{\pi^s_1} v_2(s) \geq 0. \tag{25}$$

The last inequality comes from the monotonicity of $\mathcal{T}^\pi$, which completes the proof.

Transition Invariance. From the definition of $\mathcal{T}^\pi$ in (8), we have

$$\mathcal{T}^\pi(v_1 + c\mathbf{1}) = \mathcal{T}^\pi v_1 + \gamma\langle P^o_\pi, c\mathbf{1}\rangle = \mathcal{T}^\pi v_1 + \gamma c\mathbf{1}. \tag{26}$$

Then similar for $\mathcal{T}^*$, we have,

$$\mathcal{T}^*(v_1 + c\mathbf{1}) = \mathcal{T}^* v_1 + \gamma\langle P^o_\pi, c\mathbf{1}\rangle = \mathcal{T}^* v_1 + \gamma c\mathbf{1}. \tag{27}$$

Contraction. we first show the γ-contraction property of $\mathcal{T}^\pi$ by its definition in (8),

$$|[\mathcal{T}^\pi v_1](s) - [\mathcal{T}^\pi v_2](s)| = |\gamma\langle P^0_{s,\pi(s)}, v_1 - v_2\rangle| \leq \gamma\|v_1 - v_2\|_\infty. \tag{28}$$

where the last inequality comes from the distribution nature of $P^0_{s,\pi(s)}$.

Then for $\mathcal{T}^*$ and state s, we assume $v_1(s) \geq v_2(s)$ without losing generality,

$$\begin{aligned}
|[\mathcal{T}^* v_1](s) - [\mathcal{T}^* v_2)](s)| &= [\mathcal{T}_s^{\pi^s_1}]v_1(s) - [\mathcal{T}_s^{\pi^s_2}]v_2(s) \\
&\overset{\text{(i)}}{\leq} [\mathcal{T}_s^{\pi^s_1} v_1](s) - [\mathcal{T}_s^{\pi^s_1} v_2](s) && (29)\\
&\leq |[\mathcal{T}_s^{\pi^s_1} v_1](s) - [\mathcal{T}_s^{\pi^s_1} v_2](s)| && (30)\\
&\overset{\text{(ii)}}{\leq} \gamma\|v_1 - v_2\|_\infty && (31)
\end{aligned}$$

where the first equality comes from the definitions of π^s_1 and π^s_2, inequality (i) comes from the fact that $[\mathcal{T}_s^{\pi^s_2}]v_2(s) = [\max_\pi \mathcal{T}^\pi v_2](s) \geq [\mathcal{T}_s^{\pi^s_1} v_2](s)$, and inequality (ii) comes from the contraction property of $\mathcal{T}^\pi$ in (28).

Finally, We apply $\mathcal{T}^\pi$ on our objective $V^\pi_r - \lambda_t^\top V^\pi_g$, which yields

$$\begin{aligned}
[\mathcal{T}^\pi(V^\pi_r - \lambda_t^\top V^\pi_g)](s) &= (r - \lambda_t^\top g)(s, \pi(s)) + \gamma\langle P^o_{s,\pi(s)}, V^\pi_r - \lambda_t^\top V^\pi_g)\rangle && (32)\\
&\quad + \gamma \min_{P\in\mathcal{P}}\langle P_{s,\pi(s)} - P^o_{s,\pi(s)}, V^\pi_r\rangle - \gamma\lambda_t^\top \min_{P\in\mathcal{P}}\langle P_{s,\pi(s)} - P^o_{s,\pi(s)}, V^\pi_g\rangle \\
&= r(s, \pi(s)) + \gamma\langle P^o_{s,\pi(s)}, V^\pi_r\rangle + \gamma \min_{P\in\mathcal{P}}\langle P_{s,\pi(s)} - P^o_{s,\pi(s)}, V^\pi_r\rangle && (33)
\end{aligned}$$

$$
\begin{aligned}
&- \lambda_t^\top \left(g(s,\pi(s)) + \gamma \langle P^o_{s,\pi(s)}, V_g^\pi \rangle + \gamma \min_{P\in\mathcal{P}} \langle P_{s,\pi(s)} - P^o_{s,\pi(s)}, V_g^\pi \rangle \rangle \right) \\
&= r(s,\pi(s)) + \gamma \min_{P\in\mathcal{P}} \langle P_{s,\pi(s)}, V_r^\pi \rangle \qquad (34)\\
&\quad - \lambda_t^\top \left(g(s,\pi(s)) + \gamma \min_{P\in\mathcal{P}} \langle P_{s,\pi(s)}, V_g^\pi \rangle \rangle \right) \\
&= [\mathcal{T}^\pi_{r,rob} V_r^\pi](s) - \lambda_t^\top [\mathcal{T}^\pi_{g,rob} V_g^\pi](s) \qquad (35)\\
&= (V_r^\pi - \lambda_t^\top V_g^\pi)(s) \qquad (36)
\end{aligned}
$$

where the first equality comes from the definition of consistency operator $\mathcal{T}^\pi$ in (8), the second equality is a simple rearrangement, the third equality comes from the fact that $\langle P^o_{s,\pi(s)}, V_g^\pi \rangle$ is independent of $P \in \mathcal{P}$, the fourth equality comes from the definition of robust consistency operator in 2, and the last equality comes from the contraction property of robust consistency operator (Iyengar, 2005).

Combine with the fact that operator $\mathcal{T}^\pi$ is a γ-contraction, $V_r^\pi - \lambda_t^\top V_g^\pi$ is the only stationary point of $\mathcal{T}^\pi$, the proof is complete. □

A.4 Proof for Theorem 3.5

A.4.1 Preliminary

To start with, we introduce two preliminary propositions from Scherrer et al. (2015) where Proposition A.3 builds three relations, and Proposition A.1 is a direct application of these three relations. Definition A.2 is used to simplify notation in proposition A.1.

Proposition A.1 (Lemma 2, Scherrer et al. (2015))**.** *Consider approximate modified policy iteration scheme with standard bellman operator $\mathcal{T}^\pi_{\text{st}}$.*

$$
\begin{cases}
\pi^{k+1} &= \operatorname{argmax}_\pi^{\epsilon'_{k+1}} \mathcal{T}^\pi_{\text{st}} v^k = \mathcal{G}^{\epsilon'_{k+1}} v^k \\
v^{k+1} &= (\mathcal{T}^{\pi^{k+1}}_{\text{st}})^m v^k + \epsilon_{k+1}
\end{cases} \qquad (37)
$$

where π^{k+1} is the greedy policy with respect to v^k with some error ϵ'_{k+1}, e.g. $\forall \pi \in \Pi, \mathcal{T}^{\pi^{k+1}}_{\text{st}} v^k + \epsilon'_{k+1} \geq \mathcal{T}^\pi_{\text{st}} v^k$.

Let v_* denote the optimal value function, $d_k := v_* - (\mathcal{T}^{\pi^k}_{\text{st}})^m v^{k-1}$, $s_k := (\mathcal{T}^{\pi^k}_{\text{st}})^m v^{k-1} - v_{\pi^k}$ and $b_k := v^k - \mathcal{T}^{\pi^{k+1}}_{\text{st}} v^k$, then for $k \geq 1$, we have,

$$
b_k \leq (\gamma P_{\pi^k})^m b_{k-1} + x_k \qquad (38)
$$

$$
d_{k+1} \leq \gamma P_{\pi_*} d_k + y_k + \sum_{j=1}^{m-1} (\gamma P_{\pi^{k+1}})^j b_k \qquad (39)
$$

$$
s_k = (\gamma P_{\pi^k})^m (I - \gamma P_{\pi^k})^{-1} b_{k-1} \qquad (40)
$$

where $x_k := (I - \gamma P_{\pi^k})\epsilon_k + \epsilon'_{k+1}$ and $y_k := -\gamma P_{\pi_*}\epsilon_k + \epsilon'_{k+1}$.

Definition A.2 (Γ-matrix (Scherrer et al., 2015))**.** *For a positive integer n, we define $\mathbb{P}_n$ as the smallest set of discounted transition kernels that are defined as (i) for any set of n policies $\{\pi^i\}$, $(\gamma P^o_{\pi^1})(\gamma P^o_{\pi^2}) \cdots (\gamma P^o_{\pi^n}) \in \mathbb{P}_n$, where $P^o_\pi(s,s') = P^o_{s,\pi(s)}(s')$. (ii) for any $\alpha \in (0,1)$ and $(P_1, P_2) \in \mathbb{P}_n \times \mathbb{P}_n$, $\alpha P_1 + (1-\alpha) P_2 \in \mathbb{P}_n$. With slight abuse of notation, Γ^n is used to denote any element of $\mathbb{P}_n$.*

Proposition A.3 (Lemma 4, Scherrer et al. (2015))**.** *After k iterations of approximate modified policy iteration scheme with standard bellman operator $\mathcal{T}^\pi_{\text{st}}$, the losses $l_k := v_* - v_{\pi^k}$ satisfy*

$$
l_k \leq 2 \sum_{i=1}^{k-1} \sum_{j=i}^{\infty} \Gamma^j |\epsilon_{k-i}| + \sum_{i=0}^{k-1} \sum_{j=i}^{\infty} \Gamma^j |\epsilon'_{k-i}| + h(k) \qquad (41)
$$

where $h(k) := 2\sum_{j=k}^\infty \Gamma^j |d_0|$ or $h(k) := 2\sum_{j=k}^\infty \Gamma^j |b_0|$.

A.4.2 Proof pipelne of Theorem 3.5

We first prove a technical result for our losses $l_k \coloneqq v_{\pi_t} - v_{\pi^k}$ that is similar to Proposition A.3.

Lemma A.4. *Under assumption 3.4, after k iterations of scheme (9), our losses satisfy*

$$l_k \leq 2\sum_{i=1}^{k-1}\sum_{j=i}^{\infty}\Gamma^j|\epsilon_{k-i}| + \sum_{i=0}^{k-1}\sum_{j=i}^{\infty}\Gamma^j|\epsilon'_{k-i}| + h(k), \tag{42}$$

where $h(k) \coloneqq 2\sum_{j=k}^{\infty}\Gamma^j|l_0|$ or $h(k) \coloneqq 2\sum_{j=k}^{\infty}\Gamma^j|b_0|$, with $b_0 = \mathcal{T}^{\pi^1}v^0 - v^0$ that is related to the choice of the starting point.

Proof. Similar to the proofs in Scherrer et al. (2015) for Proposition A.3, we first derive three relations that are identical to those in Proposition A.1, then apply these relations to bound our losses $l_k \coloneqq v_{\pi_t} - v_{\pi^k}$.

To start with, we define

$$b_k \coloneqq v^k - \mathcal{T}^{\pi^{k+1}}v^k, \quad s_k \coloneqq (\mathcal{T}^{\pi^k})^m v^{k-1} - v_{\pi^k}, \quad d_k \coloneqq v_{\pi_t} - (\mathcal{T}^{\pi^k})^m v^{k-1}. \tag{43}$$

Bounding b_k With the definitions of ϵ_k and ϵ'_k, and the property that

$$(\mathcal{T}^{\pi})^m v - (\mathcal{T}^{\pi})^m v' = (\gamma P^o_{\pi})^m (v - v'), \tag{44}$$

we have,

$$\begin{aligned}
b_k &= v^k - \mathcal{T}^{\pi^k}v^k + \mathcal{T}^{\pi^k}v^k - \mathcal{T}^{\pi^{k+1}}v^k && (45)\\
&\overset{\text{(i)}}{\leq} v^k - \mathcal{T}^{\pi^k}v^k + \epsilon'_{k+1} && (46)\\
&= v^k - \epsilon_k - \mathcal{T}^{\pi^k}v^k + P^o_{\pi^k}\epsilon_k + \epsilon_k - P^o_{\pi^k}\epsilon_k + \epsilon'_{k+1} && (47)\\
&\overset{\text{(ii)}}{=} v^k - \epsilon_k - \mathcal{T}^{\pi^k}(v^k - \epsilon_k) + (I - \gamma P^o_{\pi^k})\epsilon_k + \epsilon'_{k+1} && (48)\\
&= v^k - \epsilon_k - \mathcal{T}^{\pi^k}(v^k - \epsilon_k) + x_k, && (49)
\end{aligned}$$

where $x_k \coloneqq (I - \gamma P^o_{\pi^k})\epsilon_k + \epsilon'_{k+1}$, inequality (i) comes from the definition of ϵ'_{k+1} in (9), equality (ii) comes from property (44).

$$\begin{aligned}
b_k &\leq v^k - \epsilon_k - \mathcal{T}^{\pi^k}(v^k - \epsilon_k) + x_k && (50)\\
&\overset{\text{(i)}}{=} (\mathcal{T}^{\pi^k})^m v^{k-1} - \mathcal{T}^{\pi^k}(\mathcal{T}^{\pi^k})^m v^{k-1} + x_k && (51)\\
&= (\mathcal{T}^{\pi^k})^m v^{k-1} - (\mathcal{T}^{\pi^k})^m(\mathcal{T}^{\pi^k}v^{k-1}) + x_k && (52)\\
&\overset{\text{(ii)}}{\leq} (\gamma P^o_{\pi^k})^m(v^{k-1} - \mathcal{T}^{\pi^k}v^{k-1}) + x_k && (53)\\
&= (\gamma P^o_{\pi^k})^m b_{k-1} + x_k, && (54)
\end{aligned}$$

where equality (i) comes from the definition of ϵ in (9), and equality (ii) comes from property (44).

Bounding s_k With the property that $\forall v, v_{\pi^k} = (\mathcal{T}^{\pi^k})^{\infty}v$, we have

$$\begin{aligned}
s_k &= (\mathcal{T}^{\pi^k})^m v^{k-1} - (\mathcal{T}^{\pi^k})^{\infty}v^{k-1} && (55)\\
&= (\gamma P^o_{\pi^k})^m (I - \gamma P^o_{\pi^k})^{-1} b_{k-1}, && (56)
\end{aligned}$$

where the last equality comes from property (44).

Bounding d_k Define $y_k \coloneqq -\gamma P^o_{\pi_t}\epsilon_k + \epsilon'_{k+1}$, then

$$
\begin{aligned}
d_{k+1} &= v_{\pi_t} - (\mathcal{T}^{\pi^{k+1}})^m v^k && (57)\\
&\overset{\text{(i)}}{=} \mathcal{T}^{\pi_t} v_{\pi_t} - \mathcal{T}^{\pi_t} v^k + \mathcal{T}^{\pi_t} v^k - \mathcal{T}^{\pi^{k+1}} v^k + \mathcal{T}^{\pi^{k+1}} v^k - (\mathcal{T}^{\pi^{k+1}})^m v^k && (58)\\
&\overset{\text{(i)}}{\leq} \gamma P^o_{\pi_t}(v_{\pi_t} - v_k) + \epsilon'_{k+1} + \mathcal{T}^{\pi^{k+1}} v^k - (\mathcal{T}^{\pi^{k+1}})^m v^k && (59)\\
&= \gamma P^o_{\pi_t}(v_{\pi_t} - v_k) + \gamma P^o_{\pi_t}\epsilon_k - \gamma P^o_{\pi_t}\epsilon_k + \epsilon'_{k+1} + \mathcal{T}^{\pi^{k+1}} v^k - (\mathcal{T}^{\pi^{k+1}})^m v^k && (60)\\
&= \gamma P^o_{\pi_t} d_k + y_k + \mathcal{T}^{\pi^{k+1}} v^k - (\mathcal{T}^{\pi^{k+1}})^m v^k && (61)\\
&\overset{\text{(iii)}}{=} \gamma P^o_{\pi_t} d_k + y_k + \sum_{j=1}^{m-1} (\gamma P^o_{\pi^{k+1}})^j b_k, && (62)
\end{aligned}
$$

where equality (i) comes from $v_{\pi_t} = \mathcal{T}^{\pi_t} v_{\pi_t}$, inequality (ii) comes from $\mathcal{T}^{\pi_t} v_{\pi_t} - \mathcal{T}^{\pi_t} v^k = \gamma P^o_{\pi_t}(v_{\pi_t} - v_k)$ and the definition of ϵ'_{k+1}, and equality (iii) comes from iteratively applying property of $(\mathcal{T}^{\pi})^j v - (\mathcal{T}^{\pi})^j v' = (\gamma P^o_{\pi})^j (v - v')$.

After bounding all three elements, using the notation introduced in Definition A.2, we may rewrite (56) as

$$
s_k = \Gamma^m \sum_{j=0}^{\infty} \Gamma^j b_{k-1}, \tag{63}
$$

And by induction from Eq.(54) and (62), we obtain

$$
b_k \leq \sum_{i=1}^{k} \Gamma^{m(k-i)} x_i + \Gamma^{mk} b_0, \tag{64}
$$

$$
d_k \leq \sum_{j=1}^{k-1} \Gamma^{k-1-j} (y_j + \sum_{l=1}^{m-1} \Gamma^l b_j) + \Gamma^k d_0. \tag{65}
$$

Combine with the fact of $l_k = s_k + d_k$, now we recover exactl the same expression of s_k, b_k, and d_k as in the proof of proposition A.3. All the rest are therefore standard as in theirs except for the relation between b_0 and d_0,

$$
\begin{aligned}
b_0 &= v^0 - \mathcal{T}^{\pi^1} v^0 && (66)\\
&= v^0 - v_{\pi_t} + \mathcal{T}^{\pi_t} v_{\pi_t} - \pi_{\sqcup} v^0 + \pi_{\sqcup} v^0 - \mathcal{T}^{\pi^1} v^0 && (67)\\
&\leq (I - \gamma P^o_{\pi_t})(-d_0) + \epsilon'_1, && (68)
\end{aligned}
$$

where the fact of $\epsilon_0 = 0$ is used and we again recover the same relation. □

Finally we finish the proof of theorem 3.5 by noticing that $\Gamma^n \leq \gamma^n$ in Definition A.2. Therefore, we have

$$
\begin{aligned}
l_k &\leq 2 \sum_{i=1}^{k-1} \sum_{j=i}^{\infty} \Gamma^j |\epsilon_{k-i}| + \sum_{i=0}^{k-1} \sum_{j=i}^{\infty} \Gamma^j |\epsilon'_{k-i}| + h(k) && (69)\\
&\leq \frac{2\bar{\epsilon}(\gamma - \gamma^k) + \bar{\epsilon}'(1 - \gamma^k)}{(1-\gamma)^2} + O(\gamma^k) && (70)\\
&\xrightarrow{k\to\infty} \frac{2\bar{\epsilon}\gamma + \bar{\epsilon}'}{(1-\gamma)^2}. && (71)
\end{aligned}
$$

□

B Derivation of Section 4 : DRC-RL with R-Contamination Uncertainty Sets

There exist a variety class of operators of our proposed operator $\mathcal{T}^\pi$ in (8), that keep properties as in Proposition 3.3. We present one of them and upon which derive the exact consistency operator (11) used for R-contamination uncertainty set in Section 4.

Formally, $\forall \lambda_t \in \mathbb{R}_+, s \in \mathcal{S}, v \in \mathbb{R}^{\mathcal{S}}$, we define a class of consistency operators $[\mathcal{T}_1^\pi v](s) := [\mathcal{T}^\pi v](s) - \gamma h \langle P^o_{s,\pi(s)}, v - V_r^\pi + \lambda_t^\top V_g^\pi \rangle$ controlled by coefficient $h \in [0,1)$. We use P^o_π to denote vector $[P^o_{s,\pi(s)}]^\top_{s\in\mathcal{S}}$.

Proposition B.1. *The consistency operator $\mathcal{T}_1^\pi$ satisfies the following :*

(1) Monotonicity: let $v_1, v_2 \in \mathbb{R}^{\mathcal{S}}$ such that $v_1 \geq v_2$, then $\mathcal{T}_1^\pi v_1 \geq \mathcal{T}_1^\pi v_2$.

(2) Transition Invariance: for any $c \in \mathbb{R}$, we have $\mathcal{T}_1^\pi(v + c\mathbf{1}) = \mathcal{T}_1^\pi v + \gamma(1-h)c\mathbf{1}$.

(3) Contraction: The operator $\mathcal{T}_1^\pi$ is a $\gamma(1-h)$-contraction. Further, $V_r^\pi - \lambda_t^\top V_g^\pi$ is the unique stationary point.

Proof. **Monotonicity**. By the definition of $\mathcal{T}_1^\pi$, we have $\forall v_1, v_2 \in \mathbb{R}^{\mathcal{S}}$ such that $v_1 \geq v_2$,

$$\mathcal{T}_1^\pi v_1 - \mathcal{T}_1^\pi v_2 = \gamma(1-h)\langle P^o_\pi, v_1 - v_2 \rangle \geq 0. \tag{72}$$

Transition Invariance. By the definition of $\mathcal{T}_1^\pi$, we have $\forall v_1 \in \mathbb{R}^{\mathcal{S}}, c \in \mathbb{R}$

$$\mathcal{T}_1^\pi(v_1 + c\mathbf{1}) = \mathcal{T}_1^\pi v_1 + \gamma(1-h)\langle P^o_\pi, c\mathbf{1} \rangle = \mathcal{T}_1^\pi v_1 + \gamma(1-h)c\mathbf{1}. \tag{73}$$

Contraction. We first show that $\mathcal{T}_1^\pi$ is a $\gamma(1-h)$ contraction by its definition, that we have $\forall v_1, v_2 \in \mathbb{R}^{\mathcal{S}}, s \in \mathcal{S}$,

$$|[\mathcal{T}_1^\pi v_1](s) - [\mathcal{T}_1^\pi v_2](s)| = |\gamma(1-h)\langle P^0_{s,\pi(s)}, v_1 - v_2 \rangle| \leq \gamma(1-h)\|v_1 - v_2\|_\infty. \tag{74}$$

Finally, we apply $\mathcal{T}_1^\pi$ on $V_r^\pi - \lambda_t^\top V_g^\pi$,

$$\begin{aligned}
[\mathcal{T}_1^\pi(V_r^\pi - \lambda_t^\top V_g^\pi)](s) &= [\mathcal{T}^\pi(V_r^\pi - \lambda_t^\top V_g^\pi)](s) - \gamma h \langle P^o_{s,\pi(s)}, V_r^\pi - \lambda_t^\top V_g^\pi - V_r^\pi + \lambda_t^\top V_g^\pi \rangle && (75)\\
&= [\mathcal{T}^\pi(V_r^\pi - \lambda_t^\top V_g^\pi)](s) && (76)\\
&= V_r^\pi(s) - \lambda_t^\top V_g^\pi(s), && (77)
\end{aligned}$$

where the first equality comes from the definition of $\mathcal{T}_1^\pi$, the last equality comes from the contraction property of $\mathcal{T}^\pi$ in Proposition 3.3. □

By setting $h = \beta$ and specifying the uncertainty set as R-contamination uncertainty set $\mathcal{P}_{s,a} = \{(1-\beta)P^o_{s,a} + \beta q \,|\, q \in \triangle(\mathcal{S})\}$ with a scalar robust level $\beta \in \mathbb{R}$, we have for $\mathcal{T}_1^\pi$ defined in Proposition B.1 and $s \in \mathcal{S}, v \in \mathbb{R}^{\mathcal{S}}$,

$$\begin{aligned}
[\mathcal{T}_1^\pi v](s) &= [\mathcal{T}^\pi v](s) - \gamma\beta\langle P^o_{s,\pi(s)}, v - V_r^\pi + \lambda_t^\top V_g^\pi \rangle && (78)\\
&\overset{\text{(i)}}{=} (r - \lambda_t^\top g)(s,\pi(s)) + \gamma\langle P^o_{s,\pi(s)}, v\rangle + \gamma \min_{P\in\mathcal{P}} \langle P_{s,\pi(s)} - P^o_{s,\pi(s)}, V_r^\pi \rangle && (79)\\
&\quad - \gamma\lambda_t^\top \min_{P\in\mathcal{P}} \langle P_{s,\pi(s)} - P^o_{s,\pi(s)}, V_g^\pi \rangle - \gamma\beta\langle P^o_{s,\pi(s)}, v - V_r^\pi + \lambda_t^\top V_g^\pi \rangle \\
&\overset{\text{(ii)}}{=} (r - \lambda_t^\top g)(s,\pi(s)) + \gamma\langle P^o_{s,\pi(s)}, v\rangle + \gamma \min_{q\in\Delta(\mathcal{S})} \langle \beta(q - P^o_{s,\pi(s)}), V_r^\pi \rangle && (80)\\
&\quad - \gamma\lambda_t^\top \min_{q\in\Delta(\mathcal{S})} \langle \beta(q - P^o_{s,\pi(s)}), V_g^\pi \rangle - \gamma\beta\langle P^o_{s,\pi(s)}, v - V_r^\pi + \lambda_t^\top V_g^\pi \rangle \\
&= (r - \lambda_t^\top g)(s,\pi(s)) + \gamma(1-\beta)\langle P^o_{s,\pi(s)}, v\rangle + \gamma\beta(\min_{q\in\Delta(\mathcal{S})} \langle q, V_r^\pi \rangle - \lambda_t^\top \min_{q\in\Delta(\mathcal{S})} \langle q, V_g^\pi \rangle) && (81)
\end{aligned}$$

$$= (r - \lambda_t^\top g)(s, \pi(s)) + \gamma(1-\beta)\langle P^o_{s,\pi(s)}, v\rangle + \gamma\beta(\min_{s'} V_r^\pi(s') - \lambda_t^\top \min_{s'} V_g^\pi(s')). \tag{82}$$

where equality (i) comes from the definition of consistency operator $\mathcal{T}^\pi$ in (8), and equality (ii) comes from the definition of R-contamination uncertainty set.

Now we have obtained the same consistency operator as in (11) that is derived from proposed consistency operator $\mathcal{T}^\pi$ in (8), so we use the same notation as $\mathcal{T}^\pi$ in the main text for brevity sake (and one can certainly proof everything in Section 3 as they mostly depend on these three properties, but it is no need to do that as following).

In the following content of Section 4, we will further simplify to obtain a standard consistency operator, thus the convergence result is standard from Scherrer et al. (2015), such as in Proposition A.3.

C Proof of Section 5: On the Intractability of DRC-RL

C.1 Proof of Lemma 5.3

For brevity, we denote policy as $\pi \in \Delta(\mathcal{A})^{\mathcal{S}} \subseteq \mathbb{R}^{\mathcal{A}\times\mathcal{S}}$ which includes both the class of stochastic and deterministic policies. We therefore use $\pi[s,a]$ to denote the probability of action a given state s, and use $\pi(s)$ to denote the whole probability simplex given state s.

When an operator is linear, by the definition of linearity, there exists a function $f : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, such that for every value function $v \in \mathbb{R}^{|\mathcal{S}|}$ and state $s \in \mathcal{S}$,

$$\begin{aligned}
\max_{\pi\in\Pi}[\mathcal{T}^\pi v](s) &= \max_{\pi\in\Pi}\langle \pi, f(v)\rangle_s && (83)\\
&= \max_{\pi\in\Pi}\langle \pi[s,\cdot], f(v)[s,\cdot]\rangle && (84)\\
&= \max_{\pi(s)\in\Delta(\mathcal{A})}\langle \pi(s), f(v)[s,\cdot]\rangle && (85)\\
&= \max_a f(v)[s,a] && (86)
\end{aligned}$$

The greedy policy can be set as $g(v,a)_s = f(v)[s,a]$, which completes the proof for the first part.

When an operator $\mathcal{T}^\pi$ enables the greedy policy, by its definition, there exists a function $g : \mathbb{R}^{|\mathcal{S}|} \times \mathcal{A} \to \mathbb{R}^{|\mathcal{S}|}$ such that $\forall v \in \mathbb{R}^{|\mathcal{S}|}, s \in \mathcal{S}$, $\max_{\pi\in\Pi}[\mathcal{T}^\pi v](s) = \max_{a\in\mathcal{A}} g(v,a)_s$. Therefore, when defining the linear operator $\mathcal{T}^\pi_{linear} := \langle \pi(s), g(v,\cdot)\rangle$, we complete the proof by showing:

$$\begin{aligned}
\max_{\pi\in\Pi}[\mathcal{T}^\pi v](s) &= \max_{a\in\mathcal{A}} g(v,a)_s && (87)\\
&= \max_{\pi(s)\in\Delta(\mathcal{A})}\langle \pi(s), g(v,\cdot)\rangle && (88)\\
&= \max_{\pi\in\Pi}\langle \pi(s), g(v,\cdot)\rangle && (89)\\
&= \max_{\pi\in\Pi}[\mathcal{T}^\pi_{linear} v](s). && (90)
\end{aligned}$$

□

C.2 Proof for Theorem 5.4

Throughout this proof, for convenience, we denote $v^\pi := V_r^\pi - \lambda_t^T V_g^\pi$ for any policy $\pi \in \Pi$ and any λ_t.

To start with, we suppose that there exists a consistency operator denoted as $\mathcal{T}'$ that satisfies both conditions in Theorem 5.4 for all policy π, namely, $\forall s \in \mathcal{S}, \pi_1, \pi_2 \in \Pi$, we have:

$$|[\mathcal{T}'^{\pi_1} v^{\pi_1}](s) - [\mathcal{T}'^{\pi_2} v^{\pi_2}](s)| = |[\mathcal{T}'^{\pi_1} v^{\pi_1}](s) - [\mathcal{T}'^{\pi_1} v^{\pi_2}](s) + [\mathcal{T}'^{\pi_1} v^{\pi_2}](s) - [\mathcal{T}'^{\pi_2} v^{\pi_2}](s)| \tag{91}$$

$$\overset{\text{(i)}}{\leq} |[\mathcal{T}'^{\pi_1} v^{\pi_1}](s) - [\mathcal{T}'^{\pi_1} v^{\pi_2}](s)| + |[\mathcal{T}'^{\pi_1} v^{\pi_2}](s) - [\mathcal{T}'^{\pi_2} v^{\pi_2}](s)| \tag{92}$$

$$\overset{\text{(ii)}}{\leq} \gamma \left\| v^{\pi_1} - v^{\pi_2} \right\|_\infty + |\left\langle \pi_1(s) - \pi_2(s), f(v^{\pi_2})[s, \cdot] \right\rangle|, \tag{93}$$

where inequality (i) uses the triangle inequality, inequality (ii) comes from the contraction property in Proposition 5.4 and the linearity property defined in definition 5.2.

Moreover, as v^π is the stationary point of consistency operator $\mathcal{T}'^\pi$, we arrive at an inequality that is independent of any operator,

$$|v^{\pi_1}(s) - v^{\pi_2}(s)| = |[\mathcal{T}'^{\pi_1} v^{\pi_1}](s) - [\mathcal{T}'^{\pi_2} v^{\pi_2}](s)| \tag{94}$$

$$\leq \gamma \left\| v^{\pi_1} - v^{\pi_2} \right\|_\infty + |\left\langle \pi_1(s) - \pi_2(s), f(v^{\pi_2})[s, \cdot] \right\rangle|, \tag{95}$$

where the last inequality holds by applying (93).

For the rest of the proof, we will construct an example that contradicts with (95), which shows that there is no such operator $\mathcal{T}'$ that satisfies two conditions introduced in Theorem 5.4.

Figure 2: A two states, two actions Markov decision process used in example C.1 : the left and the right figures present the transition probabilities for actions a_0 and a_1.

Example C.1. *Consider Markov decision process as in figure 2, where state space $\mathcal{S} = \{s_0, s_1\}$, action space $\mathcal{A} = \{a_0, a_1\}$, discount factor $\gamma = 0.95$. The transition kernel for state s_0 and action a_0 is parameterized by ξ that indicates the probability to state s_1, i.e. $\mathcal{P}_{s_0,a_0} = \{[1-\xi, \xi] | \xi \in [0.9, 1]\}$. For other state action pairs, the only consequence is staye at the current state, i.e. $\mathcal{P}_{s_1,a_0} = \mathcal{P}_{s_1,a_1} = \{[0, 1]\}, \mathcal{P}_{s_0,a_1} = \{[1, 0]\}$.*

We assume there is one additional constraint, and design the rewards r and g for the objective and the constraint respectively,

$$r_{s,a} = \begin{cases} 1, & \text{if } s = s_0, a = a_0 \\ 0, & \text{if } s = s_0, a = a_1 \\ 1, & \text{if } s = s_1, a = a_0 \\ 2, & \text{if } s = s_1, a = a_1 \end{cases} \qquad g_{s,a} = \begin{cases} 1, & \text{if } s = s_0, a = a_0 \\ 0, & \text{if } s = s_0, a = a_1 \\ 0, & \text{if } s = s_1, a = a_0 \\ 1, & \text{if } s = s_1, a = a_1 \end{cases} \tag{96}$$

The derived inequality (95) should hold for any choice of policies π_1 and π_2, we thus choose

$$\pi_1(s) = \begin{cases} a_0, & \text{if } s = s_0 \\ a_0, & \text{if } s = s_1 \end{cases} \qquad \pi_2(s) = \begin{cases} a_0, & \text{if } s = s_0 \\ a_1, & \text{if } s = s_1 \end{cases} \tag{97}$$

The derived inequality (95) then becomes,

$$|v^{\pi_1}(s_0) - v^{\pi_2}(s_0)| \leq \gamma \left\| v^{\pi_1} - v^{\pi_2} \right\|_\infty + |\left\langle \pi_1(s_0) - \pi_2(s_0), f(v^{\pi_2})[s_0, \cdot] \right\rangle| \tag{98}$$

$$= \gamma \left\| v^{\pi_1} - v^{\pi_2} \right\|_\infty \tag{99}$$

where the second equation comes from the fact that $\pi_1(s_0) - \pi_2(s_0) = 0$.

One can then calculate the robust value function $V_r^{\pi_1}$,

$$V_r^{\pi_1}(s_1) = \frac{r_{s_1,\pi_1(s_1)}}{1-\gamma} = 20 \tag{100}$$

$$V_r^{\pi_1}(s_0) = [\mathcal{T}_{r,\text{rob}}^{\pi_1} V_r^{\pi_1}](s_0) \tag{101}$$
$$= r_{s_0,\pi_1(s_0)} + \gamma \min_{P \in \mathcal{P}_{s_0,\pi_1(s_0)}} \langle P, V_r^{\pi_1} \rangle \tag{102}$$
$$= 1 + \gamma \min_{\xi \in [0.9,1]} \{(1-\xi) V_r^{\pi_1}(s_0) + \xi V_r^{\pi_1}(s_1)\} \tag{103}$$
$$= 1 + \gamma \min_{\xi \in [0.9,1]} \{(1-\xi) V_r^{\pi_1}(s_0) + 20\xi\} \tag{104}$$
$$= 1 + \gamma \begin{cases} 20, & V_r^{\pi_1}(s_0) \geq 20 \\ 0.1 V_r^{\pi_1}(s_0) + 18, & V_r^{\pi_1}(s_0) < 20 \end{cases} \tag{105}$$

where the first two equations in calculating $V_r^{\pi_1}(s_0)$ comes from the fact that $V_r^{\pi_1}$ is the stationary point of consistency operator $\mathcal{T}_{r,\text{rob}}^{\pi_1}$ in (2), and the last equation illustrates the solution of the inner minimization problem in (104).

One can easily calculate both cases in (105) and find that $V_r^{\pi_1}(s_0) = 20$. Similarly, we can calculate other robust value functions,

$$V_g^{\pi_1}(s_1) = \frac{g_{s_1,\pi_1(s_1)}}{1-\gamma} = 0 \tag{106}$$
$$V_g^{\pi_1}(s_0) = g_{s_0,\pi_1(s_0)} + \gamma \min_{\xi \in [0.9,1]} \{(1-\xi) V_g^{\pi_1}(s_0) + \xi V_g^{\pi_1}(s_1)\} = 1 \tag{107}$$
$$V_r^{\pi_2}(s_1) = \frac{r_{s_1,\pi_2(s_1)}}{1-\gamma} = 40 \tag{108}$$
$$V_r^{\pi_2}(s_0) = r_{s_0,\pi_2(s_0)} + \gamma \min_{\xi \in [0.9,1]} \{(1-\xi) V_r^{\pi_2}(s_0) + \xi V_r^{\pi_2}(s_1)\} \tag{109}$$
$$= 0 + \gamma \begin{cases} 40, & V_r^{\pi_1}(s_0) \geq 40 \\ 0.1 V_r^{\pi_1}(s_0) + 36, & V_r^{\pi_1}(s_0) < 40 \end{cases} \tag{110}$$
$$= \frac{6840}{181} \approx 37.79 \tag{111}$$
$$V_g^{\pi_2}(s_1) = \frac{g_{s_1,\pi_2(s_1)}}{1-\gamma} = 20 \tag{112}$$
$$V_g^{\pi_2}(s_0) = g_{s_0,\pi_2(s_0)} + \gamma \min_{\xi \in [0.9,1]} \{(1-\xi) V_g^{\pi_2}(s_0) + \xi V_g^{\pi_2}(s_1)\} = 20 \tag{113}$$

Therefore, we have

$$v^{\pi_1} = V_r^{\pi_1} - \lambda_t V_g^{\pi_1} = \begin{bmatrix} 20 \\ 20 \end{bmatrix} - \lambda_t \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 20 - \lambda_t \\ 20 \end{bmatrix} \tag{114}$$
$$v^{\pi_2} = V_r^{\pi_2} - \lambda_t V_g^{\pi_2} = \begin{bmatrix} \frac{6840}{181} \\ 40 \end{bmatrix} - \lambda_t \begin{bmatrix} 20 \\ 20 \end{bmatrix} = \begin{bmatrix} \frac{6840}{181} - 20\lambda_t \\ 40 - 20\lambda_t \end{bmatrix} \tag{115}$$

Finally, for every $\lambda_t \in [0.969, 2.209]$, we have

$$|v^{\pi_1}(s_0) - v^{\pi_2}(s_0)| = \|v^{\pi_1} - v^{\pi_2}\|_\infty > \gamma \|v^{\pi_1} - v^{\pi_2}\|_\infty \tag{116}$$

which gives a clear contradiction to derived inequality (99).

□

Remark C.2. *The simplified derived inequality (99) is true to the non-robust or non-constrained counterparts of DRC-RL.*

D Additional Discussions

D.1 Discussions about Assumption 3.4

We start from the definitions of our consistency operator $\mathcal{T}^\pi$ in (8) and the corresponding optimality operator $\mathcal{T}^*$,

$$\begin{aligned}[\mathcal{T}^\pi v](s) &= (r - \lambda_t^\top g)(s, \pi(s)) + \gamma \langle P^o_{s,\pi(s)}, v\rangle + \gamma \min_{P\in\mathcal{P}} \langle P_{s,\pi(s)} - P^o_{s,\pi(s)}, V_r^\pi\rangle \\ &\quad - \gamma \lambda_t^\top \min_{P\in\mathcal{P}} \langle P_{s,\pi(s)} - P^o_{s,\pi(s)}, V_g^\pi\rangle \\ [\mathcal{T}^* v](s) &= \max_{\pi\in\Pi} [\mathcal{T}^\pi v](s)\end{aligned}$$

Given that consistency operator $\mathcal{T}^\pi$ does not have linearity defined in definition 5.2, the optimality operator $\mathcal{T}^*$ cannot take greedy policy as proved in lemma 5.3. Further, for a fixed value function v and different states s and s', there might not exist a single policy that can be returned by the optimality operator $\mathcal{T}^*$ acting on both states, i.e. maximize $\mathcal{T}^\pi v(s)$ and $\mathcal{T}^\pi v(s')$ at the same time. Therefore, assumption 3.4 for the policy improvement step actually consists of two statements : (1) There exists a policy that can be ϵ'-approximate to the optimality operator. (2) There exists a solver that efficiently finds this policy.

In this discussion, we will introduce a modified operator that guarantees the existence of a ϵ'-approximate policy to the optimality operator. Then further build the related AMPI scheme to solve distributionally robust constrained RL.. To start with, we first denote $v(\mu) \coloneqq \langle v, \mu\rangle$ for any vector $v \in \mathbb{R}^{\mathcal{S}}$. We then define the μ-consistency operator $\mathcal{T}_\mu^\pi : \mathbb{R} \to \mathbb{R}$,

$$\mathcal{T}_\mu^\pi v(\mu) = [\mathcal{T}^\pi v](\mu) \tag{117}$$

and its corresponding μ-optimality operator is,

$$\mathcal{T}_\mu^* v(\mu) = \max_{\pi\in\Pi} \mathcal{T}_\mu^\pi v(\mu) = \max_{\pi\in\Pi} \{[\mathcal{T}^\pi v](\mu)\} \tag{118}$$

As μ-optimality operator now gives a scalar, for any value function $v \in \mathbb{R}^{\mathcal{S}}$, there always exist a policy π_v such that $\mathcal{T}_\mu^* v(\mu) = \mathcal{T}_\mu^{\pi_v} v(\mu)$. And the existence of ϵ'-approximate policy in the policy improvement step is automatically proved, and we further have the existence of optimal policy $\pi^* = \pi_t$.

Proposition D.1. *The μ-consistency operator $\mathcal{T}_\mu^\pi$ and the optimality operator $\mathcal{T}^*$ have the following properties,*

1. *Monotonicity: let $v_1, v_2 \in \mathbb{R}^{|\mathcal{S}|}$ such that $v_1(\mu) \geq v_2(\mu)$, then $\mathcal{T}_\mu^\pi v_1(\mu) \geq \mathcal{T}_\mu^\pi v_2(\mu)$ and $\mathcal{T}_\mu^* v_1(\mu) \geq \mathcal{T}_\mu^* v_2(\mu)$.*
2. *Transition Invariance: for any $c \in \mathbb{R}$, we have $\mathcal{T}_\mu^\pi(v + c\mathbf{1})(\mu) = \mathcal{T}_\mu^\pi v(\mu) + \gamma c\mathbf{1}(\mu)$ and $\mathcal{T}_\mu^*(v + c\mathbf{1})(\mu) = \mathcal{T}_\mu^* v(\mu) + \gamma c\mathbf{1}(\mu)$*
3. *Contraction: The operator $\mathcal{T}_\mu^\pi$ and $\mathcal{T}_\mu^*$ are γ-contractions, whose unique stationary points are $(V_r^\pi - \lambda_t^\top V_g^\pi)(\mu)$ and $(V_r^{\pi_t} - \lambda_t^\top V_g^{\pi_t})(\mu)$ respectively.*

Proof. By the definition of the μ-consistency operator $\mathcal{T}_\mu^\pi v(\mu) = [\mathcal{T}^\pi v](\mu)$, it is straightforward that the monotonicity, transition invariance, and contraction properties hold for μ-consistency operator. And that the unique stationary point of μ-consistency operator is $(V_r^\pi - \lambda_t^\top V_g^\pi)(\mu)$.

We then consider μ-optimality operator $\mathcal{T}_\mu^* v(\mu)$.

Monotonicity Let π_1 satisfies $\mathcal{T}_\mu^* v_1(\mu) = \mathcal{T}_\mu^{\pi_1} v_1(\mu)$ and similar to π_2, we have

$$\mathcal{T}_\mu^* v_1(\mu) - \mathcal{T}_\mu^* v_2(\mu) = \max_\pi \mathcal{T}_\mu^\pi v_1(\mu) - \mathcal{T}_\mu^{\pi_2} v_2(\mu) \geq \mathcal{T}_\mu^{\pi_2} v_1(\mu) - \mathcal{T}_\mu^{\pi_2} v_2(\mu) \geq 0 \tag{119}$$

Transition Invariance

$$\mathcal{T}_\mu^*(v_1 + c\mathbf{1})(\mu) = \max_\pi \mathcal{T}_\mu^\pi(v_1 + c\mathbf{1})(\mu) = \max_\pi \mathcal{T}_\mu^\pi v_1(\mu) + \gamma c\mathbf{1}(\mu) \tag{120}$$

Contraction We first assume $\mathcal{T}_\mu^* v_1(\mu) > \mathcal{T}_\mu^* v_2(\mu)$ without loss of generality,

$$|\mathcal{T}_\mu^* v_1(\mu) - \mathcal{T}_\mu^* v_2(\mu)| = \mathcal{T}_\mu^{\pi_1} v_1(\mu) - \max_\pi \mathcal{T}_\mu^\pi v_2(\mu) \leq \mathcal{T}_\mu^{\pi_1} v_1(\mu) - \mathcal{T}_\mu^{\pi_1} v_2(\mu) \leq \gamma |v_1(\mu) - v_2(\mu)| \tag{121}$$

Then as $(V_r^{\pi_t} - \lambda_t^\top V_g^{\pi_t})(\mu)$ is the unique stationary point of $\mathcal{T}_\mu^{\pi_t}$, we have $(V_r^{\pi_t} - \lambda_t^\top V_g^{\pi_t})(\mu) \leq \mathcal{T}_\mu^{\pi^*}(V_r^{\pi_t} - \lambda_t^\top V_g^{\pi_t})(\mu) \leq (V_r^{\pi^*} - \lambda_t^\top V_g^{\pi^*})(\mu)$ where π^* is the stationary policy of $\mathcal{T}_\mu^{\pi^*}$.

At the same time, by the definition of π_t, we have,

$$(V_r^{\pi_t} - \lambda_t^\top V_g^{\pi_t})(\mu) \geq (V_r^{\pi^*} - \lambda_t^\top V_g^{\pi^*})(\mu) \tag{122}$$

Therefore, $(V_r^{\pi_t} - \lambda_t^\top V_g^{\pi_t})(\mu)$ is the unique stationary point of $\mathcal{T}_\mu^{\pi^*}$. □

Now consider the related AMPI scheme, with which the assumption 3.4 only assumes the availability of approximate solvers,

$$\begin{cases} \pi^{k+1} &= \operatorname{argmax}_\pi^{\epsilon'_{k+1,\mu}} \mathcal{T}_\mu^\pi v^k(\mu) \\ v^{k+1} &= (\mathcal{T}^{\pi^{k+1}})^m v^k + \epsilon_{k+1} \end{cases} \tag{123}$$

Unlike AMPI scheme (9), the error for policy improvement step $\epsilon'_\mu \in \mathbb{R}$ is a scalar that guarantees $\max_\pi \mathcal{T}_\mu^\pi v^k(\mu) \leq \mathcal{T}_\mu^{\pi^{k+1}} v^k(\mu) + \epsilon'_{k+1,\mu}$. We then define the non-negative scalar loss $l_{k,\mu} = v_{\pi_t}(\mu) - v_{\pi^k}(\mu)$ and finally obtain its absolute error bounds.

Theorem D.2. *Under assumption 3.4, after k iterations of scheme (123), the losses satisfy*

$$|l_{k,\mu}| \leq 2\sum_{i=1}^{k-1}\sum_{j=i}^{\infty} \Gamma^j |\epsilon_{k-i}(\mu)| + \sum_{i=0}^{k-1}\sum_{j=i}^{\infty} \Gamma^j |\epsilon'_{k-i,\mu}| + h(k) \tag{124}$$

where $h(k) \coloneqq 2\sum_{j=k}^\infty \Gamma^j |l_{0,\mu}|$ or $h(k) \coloneqq 2\sum_{j=k}^\infty \Gamma^j |b_{0,\mu}|$, with $b_{0,\mu} = (\mathcal{T}^{\pi^1} v^0 - v^0)(\mu)$ that is related to the choice of the starting point.

Proof. Given the relation of $\mathcal{T}_\mu^\pi v(\mu) = [\mathcal{T}^\pi v](\mu)$ and the properties in proposition D.1, one can define $b_{k,\mu} = b_k(\mu)$ and similar for $s_{k,\mu}$ and $d_{k,\mu}$. Similar to proof for theorem 3.5, we have

$$b_{k,\mu} \leq \sum_{i=1}^{k} \Gamma^{m(k-i)} x_i(\mu) + \Gamma^{mk} b_{0,\mu} \tag{125}$$

$$d_{k,\mu} \leq \sum_{j=1}^{k-1} \Gamma^{k-1-j}(y_j(\mu) + \sum_{l=1}^{m-1} \Gamma^l b_{j,\mu}) + \Gamma^k d_{0,\mu} \tag{126}$$

$$s_{k,\mu} = \Gamma^m \sum_{j=0}^{\infty} \Gamma^j b_{k-1,\mu} \tag{127}$$

and the relation between $b_{0,\mu}$ and $d_{0,\mu}$,

$$b_{0,\mu} \leq (I - \gamma P_{\pi_t}^o)(-d_{0,\mu}) + \epsilon'_{1,\mu} \tag{128}$$

Then follow with the proof in proposition A.3, we complete by noticing $l_{k,\mu}$ is non-negative given the definition of `Best-response`π_t. □

Algorithm 2 DRC-RL with Contamination Set

Input : Dataset $\mathcal{D} = \{x_i, a_i, x_i', r_i, g_i\}_{i=1}^n$. Online algorithm parameters: ℓ_1 norm bound B, learning rate η. Robust level β. Initial discounted horizon γ_0.

Initialize $\lambda_1 = (\frac{B}{m+1}, \ldots, \frac{B}{m+1}) \in \mathbb{R}^{m+1}$
Initialize $\gamma = (1-\beta)\gamma_0$
for each round t do **do**
  Learn $\pi_t \leftarrow \text{FQI}(r - \lambda_t^\top g)$
  Evaluate $\hat{R}(\pi_t) \leftarrow \text{FQE}(\pi_t, r), \hat{G}(\pi_t) \leftarrow \text{FQE}(\pi_t, g)$
  $\hat{\pi}_t \leftarrow \frac{1}{t}\sum_{t'=1}^t \pi_{t'}, \hat{\lambda}_t \leftarrow \frac{1}{t}\sum_{t'=1}^t \lambda_{t'}$
  $\hat{R}(\hat{\pi}_t) \leftarrow \frac{1}{t}\sum_{t'=1}^t \hat{R}(\pi_{t'})$
  $\hat{G}(\hat{\pi}_t) \leftarrow \frac{1}{t}\sum_{t'=1}^t \hat{G}(\pi_{t'})$
  Learn $\tilde{\pi} \leftarrow \text{FQI}(r - \hat{\lambda}_t^\top g)$
  Evaluate $\hat{R}(\tilde{\pi}) \leftarrow \text{FQE}(\tilde{\pi}, r), \hat{G}(\tilde{\pi})\text{FQE}(\tilde{\pi}, g)$
  $\hat{L}_{\max} = \hat{R}(\tilde{\pi}) - \hat{\lambda}_t^\top \left[(\hat{G}(\tilde{\pi}) - \tau)^\top, 0\right]^\top$
  $\hat{L}_{\min} = \min_{\lambda, \|\lambda\|_1 = B} \left(\hat{R}(\hat{\pi}_t) - \hat{\lambda}_t^\top [(\hat{G}(\hat{\pi}_t) - \tau)^\top, 0]^\top\right)$
  if $\hat{L}_{max} - \hat{L}_{min} \leq \omega$ **then**
    return $\hat{\pi}_t$
  end if
  Set $z_t \leftarrow \left[(\tau - \hat{G}(\pi_t))^\top, 0\right]^\top \in \mathbb{R}^{m+1}$
  $\lambda_{t+1}[i] \leftarrow B\frac{\lambda_t[i]e^{-\eta z_t[i]}}{\sum_j \lambda_t[j]e^{-\eta z_t[j]}} \quad \forall i$
end for

D.2 An Instantiation of the Solution with Contamination Set

In this section, we give a specific instantiation for our solution to distributionally robust constrained RL (DRC-RL) with R-contamination uncertainty sets. In short, as our solution simplified to the non-robust counterpart of DRC-RL with asmaller discount factor, we use the same subroutines as in Le et al. (2019) and get the theoretical results with generalization needs.

In specific, `Best-response` is instatiated with fitted Q iteration (FQI) (Ernst et al., 2005), `Online-Algorithm` is chosen as exponentiated gradient (EG) (Kivinen & Warmuth, 1997). The EG algorithm requires bounded λ, we thus force $\|\lambda\|_1 \leq B$. When evaluating Lagrangian, e.g. $L_{\max}$ and $L_{\min}$ in meta algorithm 1, fitted Q evaluation (FQE) (Le et al., 2019) is used. We represent the resulting instantiation in algorithm 2, which has its true performance that could be arbitrarily close to the optimal policy for DRC-RL.

Proposition D.3 (Theorem 4.4, Le et al. (2019))**.** *Let π^* be the optimal policy to (3). Denote $\bar{V} = \bar{R} + B\bar{G}$. Let K be the number of iterations of FQE and FQI, $\hat{\pi}$ be the policy returned by algorithm 2, with termination threshold ω and robust level β. For $\epsilon > 0$ and $\delta \in (0,1)$, when $n = \mathcal{O}(\frac{\bar{V}^4}{\epsilon^2}(\log\frac{K(m+1)}{\delta} + \dim_F \log\frac{\bar{V}^2}{\epsilon^2} + \log\dim_F))$, we have with probability at least $1-\delta$:*

$$V_r^{\hat{\pi}}(\mu) \geq V_r^{\pi^*}(\mu) - \omega - \frac{(4+B)(1-\beta)\gamma}{(1-(1-\beta)\gamma)^3}(\sqrt{\beta_\mu}\epsilon + 2((1-\beta)\gamma)^{K/2}\bar{V}), \tag{129}$$

and

$$V_g^{\hat{\pi}}(\mu) \geq \tau - 2\frac{\bar{V}+\omega}{B} - \frac{((1-\beta)\gamma)^{1/2}}{(1-(1-\beta)\gamma)^{3/2}}(\sqrt{\beta_\mu}\epsilon + \frac{2((1-\beta)\gamma)^{K/2}\bar{V}}{(1-(1-\beta)\gamma)^{1/2}}) \tag{130}$$

where $\dim_F$ is the pseudo-dimension (Hastie et al., 2009) for function approximation class F used in FQI and FQE , β_μ is the concentration coefficient of future state-action distribution (Munos, 2007; Le et al., 2019).

Proof. The proof is standard as in Le et al. (2019) up to a different discount factor. □

E Experiments

In this section, we illustrate the detailed setting of our verification for a solution with R-contamination uncertainty set as in Section 4. Notably, the resulting operator (12) for `Best-response` subroutine coincides with the standard consistency operator, which is used for `Best-response` subroutine in Constrained RL (Le et al., 2019), up to a discount factor difference. We thus focus on the joint constrained-robustness comparison as our target, and refer to constrained RL literature (Miryoosefi et al., 2019; Le et al., 2019) for broader comparisons.

Algorithm Design. We adopt the same choices of `Best-response` and `Online-Algorithm` with a shortened discount factor in subroutines as in Le et al. (2019). In specific, the `Best-response` algorithm is instantiated with Fitted Q Iteration (FQI) (Ernst et al., 2005), a model-free off-policy learning approach on the well-collected dataset from Le et al. (2019) with modifications on constraints. And we use multi-layered CNNs. The `Online-Algorithm` is chosen as Exponentiated Gradient (EG) algorithm (Kivinen & Warmuth, 1997), a variant of online mirror descent. The EG algorithm requires bounded λ, we thus force $\|\lambda\|_1 \leq B$ which has minor harm on the theoretical analysis as shown in proposition D.3. To meet the requirement of proposition 3.2, we use the simulator to precisely evaluate value functions of specific policies and Lagrangians, as shown in Algorithm 2. We set the initial position of our car as fixed, i.e. $\mu = \{s_0\}$, for the accuracy of all the evaluations.

E.1 Car Racing

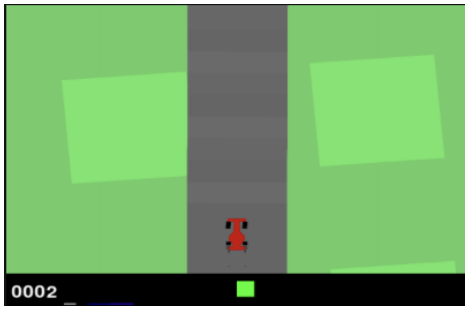


Figure 3: Car Racing environment

The environment is chosen as the car racing environment, a high-dimensional domain from Gymnasium (Towers et al., 2023), as shown in figure 3. This environment is a racetrack, where each state $s \in \mathcal{S}$ is a $96 \times 96 \times 3$ tensor of raw pixels. Given each state, the agent has 12 actions $a \in \mathcal{A} = \{(i, j, k) | i \in \{-1, 0, 1\}, j \in \{0, 1\}, k \in \{0, .2\}\}$, corresponding to steering angle, amount of gas applied and amount of brake applied, respectively. In each episode, the goal is to traverse over 95% of the track, measured by the number of tiles which amount to 281 tiles in total. The agent receives a reward of $\frac{1000}{281}$ for passing each single tile and no reward if off-track. A small positive cost of .1 applies at every time step with a maximum horizon of 1000. We further utilize the popular frame-stacking option that is common in practice in online RL for Atari and video games.

We describe the two constraints we studied as slow driving and edge driving. In slow driving constraint, the agent receives a reward $g_0(s, a) = 1$ if a contains braking action and 0 otherwise. In edge driving, the agent receives its normalized Euclidean distance of the closest point between the track and the agent from the lane center as its reward $g_1(s, a) \in [0, 10]$. Both constraints may intervene with the primary goal of track traversing. Let N_t be the number of tiles that are traversed by the agent, then the distributionally robust constrained reinforcement learning problem is :

$$\max_{\pi \in \Pi} \min_{\mathcal{K} \in \otimes_{t \geq 0} \mathcal{P}} \mathbb{E}_{\mathcal{K}}[\sum_{t=0}^{\infty} \gamma^t (\frac{1000}{281}(N_{t+1} - N_t) - .1) | s_0 = s, \pi] \tag{131}$$

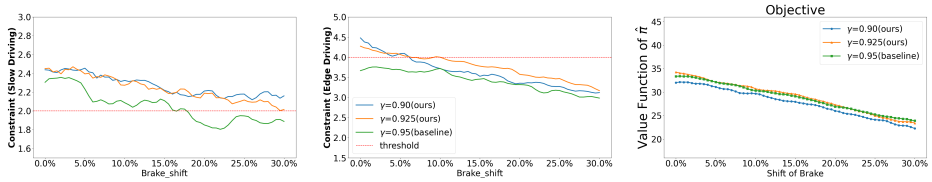
$$\text{s.t.} \quad \min_{\mathcal{K}\in\otimes_{t\geq 0}\mathcal{P}} \mathbb{E}_{\mathcal{K}}[\sum_{t=0}^{\infty}\gamma^t \mathbb{I}(a_t \in \mathcal{A}_{braking})|s_0 = s, \pi] \geq \tau_0$$

$$\min_{\mathcal{K}\in\otimes_{t\geq 0}\mathcal{P}} \mathbb{E}_{\mathcal{K}}[\sum_{t=0}^{\infty}\gamma^t \frac{10d(s_t)}{d_{max}}|s_0 = s, \pi] \geq \tau_1.$$

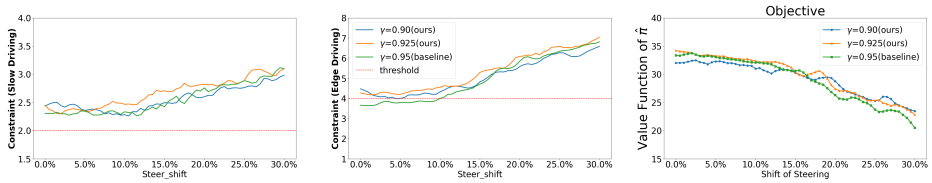
We finally set the thresholds $\tau = [\tau_0, \tau_1] = [2, 4]$ being close to the constraints satisfaction of baseline ($\gamma = 0.95$), such that the slow driving constraint is satisfied and the edge driving constraint is violated.

E.2 Full Results

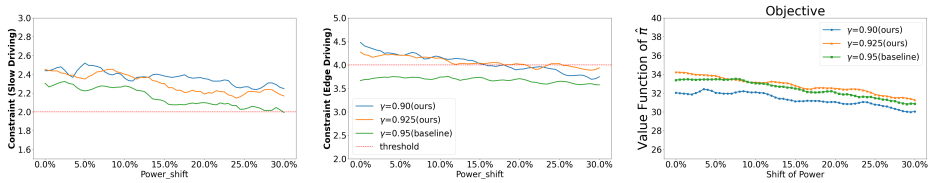
In figure E.2, the full results of our car racing experiments are presented.



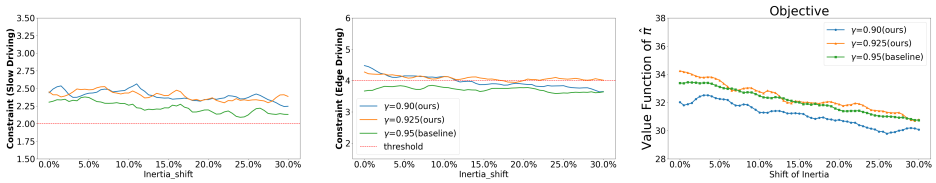
(a) Shift of Braking Magnitude



(b) Shift of Steering Angle



(c) Shift of Power



(d) Shift of Inertia of Wheel

Figure 4: Full results with Four different shifts: Higher is better, Left two are constraints and the right one is the objective. The bar graphs of constraints satisfaction are directly produced from these results.